DOCUMENT RESUME

ED 401 308 TM 025 869

AUTHOR Yang, Wen-Ling; Houang, Richard T.

TITLE The Effect of Anchor Length and Equating Method on the Accuracy of Test Equating: Comparisons of Linear and IRT-Based Equating Using an Anchor-Item Design.

PUB DATE 11 Apr 96

NOTE 84p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC04 Plus Postage.

DESCRIPTORS *Equated Scores; Estimation (Mathematics); *Guessing (Tests); *Item Response Theory; Minimum Competency Testing; *Test Format; Test Items; *True Scores

IDENTIFIERS Accuracy; *Anchor Tests; *Linear Equating Method; Tucker Common Item Equating Method

ABSTRACT

The influence of anchor length on the accuracy of test equating was studied using Tucker's linear method and two Item-Response-Theory (IRT) based methods, focusing on whether equating accuracy improved with more anchor items, whether the anchor effect depended on the equating method used, and the adequacy of the inclusion of the guessing parameter for a test that had a negatively skewed distribution of scores. Data were from 2 forms of a minimum competency examination that contained 197 and 203 items respectively. Three pairs of shorted forms were created by the simple random sampling of items, and the pairs were equated separately. The total score on the 145 anchor items was used as a criterion, a pseudo true score, to evaluate result accuracy. True score estimates were obtained that were correlated to the pseudo true score. Overall, results yielded by all three equating methods were moderately accurate, and no matter which equating method was used, the results tended to be more accurate when there were more anchor items. In addition, inclusion of a guessing parameter was justified. Six appendixes present the sampling scheme for the reduced forms, item correlations and descriptive test statistics for the reduced and full forms, and the two equating methods. (Contains 1 figure, 4 tables, 18 appendix tables, and 46 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 308

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
■ This document has been reproduced as received from the person or organization originating it.
□ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
WEN-LING YANG
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

The Effect of Anchor Length and Equating Method on the Accuracy of Test Equating: Comparisons of Linear and IRT-Based Equating Using Anchor-Item Design

by

Wen-Ling Yang and Richard T. Houang
Michigan State University

Paper Presented at the AERA Annual Conference, Division D. New York, April 11, 1996

TM025869

BEST COPY AVAILABLE

Table of Contents

Abstract ---------- 3

Introduction ---------- 5

Research Purposes ---------- 7

Research Questions and Limitation ---------- 7

Literature Review
Conditions of Equivalency ---------- 8
Same Construct
Equity
Symmetry
Population Invariance
Unidimensionality for IRT Equating

Overall Equating Guidelines ---------- 11

Criteria for Selecting Equating Methods ---------- 12
Tenability of Model Assumptions
Feasibility of Design and Method
Equating Accuracy

Tucker's Linear Equating ---------- 15

Advantages of IRT Equating ---------- 16
Curvilinear Equating
Item-Free & Person-Independent Measures
Practical Appeals

IRT Equating Methods ---------- 19
Two-stage Method
Fixed-b Method
True Score Equating
Concurrent Calibration Method
Characteristic Curve Transformation (Formula Method)
IRT Pre-equating
The Usé of IRT Equating Coefficients

Sampling Effects on Equating Results ---------- 25
Effect of Ability Difference
Representative vs. Matched Sampling

Equating Tests with Skewed Distributions ---------- 27

Assessing Equating Adequacy ---------- 28
Measures of Equating Accuracy----------28
Selection of criterion equating
Equating to self
Estimated RMSD and BIAS
Scale Stability----------31

Assumption of Unidimensionality ---------- 33
Effect of Dimensionality
Robustness of the Unidimensionality Assumption

Characteristics of Anchor items ---------- 35
Content Representativeness
Length of Anchor

Precaution: Limits of Equating ---------- 37

Description of Data ---------- 38

Content and Format of the Test ---------- 39

Examinee Groups ---------- 39

Research Design ---------- 40

Anchor-item Design of Equating ---------- 42

Random Sampling of Items ---------- 42

Equating Methods ---------- 43

Research Tool ---------- 44

Results and Discussion
Test Homogeneity & Content Representativeness ---------- 46
Preliminary Study on the Raw Scores ---------- 47
Estimation of Item Parameters ---------- 49
Equating the IRT Ability Scores ---------- 49
Estimation of the True scores ---------- 51
Tucker's Linear Equating Results ---------- 51
Accuracy of Equating ---------- 53
IRT true score estimates & Tucker's scaled scores---53
Anchor Effect----------55
IRT estimates for unique items ----------55
IRT estimates for anchor items ----------57
Critiques on the criteria of equating accuracy------58
Adequacy of the guessing parameter ----------60

Suggestions ---------- 60

Reference ---------- 63

Appendices

TABLE 1 Average Item Difficulty---------- 48
TABLE 2 Summary of IRT Calibration---------- 50
TABLE 3 Summery of Tucker's Linear Equating Results---------- 52
TABLE 4 Correlation Analysis Between "True Scores" and Various True Score Estimates---------- 54
FIGURE 1 Basic Research Design---------- 41

1996 AERA Annual Conference, Division D, Paper Presentation
Wen-Ling Yang and Richard T. Houang, Michigan State University

The Effect of Anchor Length and Equating Method on the Accuracy of Test Equating: Comparisons of Linear and IRT-Based Equating Using Anchor-Item Design

Abstract

Multiple test forms or editions have long been used to satisfy the demand of test security and to measure growth or trend. As a result, different equating methods have been frequently used to generate comparable scores from different test forms. One popular approach is to embed a set of common items in different forms. A practical issue in such design is the effects of particular equating methods and the characteristics of the anchor on the accuracy of the derived equivalent scores. This paper reports on a study that examined the influence of anchor length on the accuracy of equating. The differences between the equating results yielded by Tucker's linear method and two IRT-based methods were studied. The investigation focused on whether equating accuracy improved with more anchor items, whether the anchor effect depended on the particular equating method used, and the adequacy of the inclusion of the guessing parameter for a test that had a negatively skewed distribution of scores.

The data used was from the two forms of a minimum competency examination that contained 197 and 203 items, respectively. Equating was made possible by the 145 anchor items embedded in both of the forms. Three pairs of shorter forms were created by simple random sampling of items, with a control on the anchor length, and the pairs of the forms were equated separately. The total score on the 145 anchor items was used as a criterion, the pseudo "true score", to evaluate the accuracy of equating results. The use of such pseudo "true score" made sense because the performance of all the examinees on the "anchor universe" was attainable and the anchors in the test forms were representative subsets of the universe. However, such criterion was only appropriate when the examinee population, the test items, and the testing occasion were considered fixed. Though it was still a close approximation to the true score, the criterion was conceptually superior than the other criteria and would not be biased in favor of the IRT equating. The lower bound of IRT equating accuracy, hence, was estimated.

In addition to the Tucker's linear equating, the three parameter logistic IRT model of equating was applied. The IRT two-stage method and the IRT fixed-b's method, were used operationally to investigate the effect of IRT calibration on test equating. Lord's true score formula was applied to obtain true score estimates, which were correlated to the pseudo "true score" to estimate the equating accuracy. The PC version of BILOG 3 was used to estimate IRT item parameters and person ability.

The results of the classical item analysis showed that the items had different item difficulties and correlated moderately to the total test score. The equivalent ability scores yielded by the two IRT methods were very similar. The correlation between the two sets of scores was almost perfect, showing that the two IRT methods ordered individual examinees in an almost identical way. Regardless of the differences in anchor length, the true score estimates based on the equating results of the two methods also correlated in an expected manner, .999 over all the shortened tests. Overall, the results yielded by all three equating methods were moderately accurate. For IRT equating, the correlation coefficient between the pseudo "true scores" and the true score estimates ranged from .83 to .86 over pairs of forms. For Tucker's equating, it ranged from .80 to .83. Although the correlation coefficients did not differ much, the IRT equating always yielded more accurate scores than Tucker's linear equating.

For IRT two-stage equating, when the number of anchor items increased from 12 to 20 to 30, the correlation coefficient increased from .832 to .847 to .856. For IRT fixed-b equating, the coefficient increased from .832 to .847 to .854. And for Tucker's equating, it increased from .802 to .823 to .832. This pattern suggested that, no matter which equating method was used, the equating tended to be more accurate when there were more anchor items. However, the improvement might not be practically significant. If sufficiency and efficiency were of equal concern, it was desirable to have a minimum of 20 anchor items, or at least 1/5 of the total test should be anchor items. Since the length of the anchor might have substantial impacts on the results of equating, regardless of equating method, it was important to include a sufficient number of anchor items.

To correct for the overestimation of the equating accuracy due to auto-correlation, the pseudo "true scores" were correlated to the IRT true score estimates that based on only the non-anchor items for a partial control over the auto-correlation. Although the auto-correlation could not be completely ruled out, it provided a better picture for the goodness of equating methods. Moreover, the pseudo "true scores" were correlated with the IRT true score estimates based on the anchor items only to investigate the reliability of the anchor. The patterns of the correlation coefficients remained unchanged. The strong correlation between the criterion and the anchor provided validity and reliability evidence for the anchor. Given the valid and reliable equivalent scores, it was concluded that both IRT two-stage and fixed-b methods were satisfactory in equating the minimum competency test forms.

With the above findings, it was concluded that the three parameter IRT model fit the data used in this study. The inclusion of the guessing parameter was justified theoretically and empirically. It seemed appropriate to include the guessing parameter when equating tests that had negatively skewed score distributions.

The Effect of Anchor Length and Equating Method on the Accuracy of Test Equating: Comparisons of Linear and IRT-Based Equating Using Anchor-Item Design

Introduction

In testing situations, often not all examinees take the same test at the same occasion. To ensure test security, there is a need for alternative test forms. To measure growth or trend, interchangeable parallel forms are also needed so that test scores are comparable. The need for various test editions or forms is especially urgent for licensure exams and any other tests of which the testing results inform critical decisions. Theoretically, it is possible to obtain parallel test forms by carefully constructing a test such that the items of alternative forms have similar average difficulty and difficulty distribution. However, the result of test construction is often not satisfactory because the test forms are hardly parallel. It is hence necessary to establish equivalent scores for scores on different forms.

A variety of equating techniques have been developed to yield comparable test scores. They are primarily based on the possibility of making statistical adjustment to approach testing equivalency. From the perspective of transforming test scores across forms, equating can be linear or non-linear. Equating can also be categorized into classical linear equating or item response theory (IRT) application, of which the assumptions, mathematical functions, and computational procedures are substantially different.

The selection of an equating model depends on the purpose of equating, the underlying theory of equating, the feasibility and accuracy of the model, as well as the characteristics of examinees and test data. Classical linear equating methods have being popular for years for their straightforward conceptual steps and convenient computations. However, the equating results sometimes do not meet all the needs. For example, the item calibration varies across examinee groups and item samples. To overcome the drawbacks of classical equating, equating models derived from item response theory are increasingly applied to large-scale testing. Nevertheless, there are still doubts about the accuracy of IRT equating, its practical value, and the claimed superiority over classical equating methods.

With respect to the need of better equating and the controversies in practice, the current study sought to resolve some important issues in both theoretical and empirical ways. A comprehensive literature review of the underlying theories for test equating and their applications was included. Pairs of test forms, varying in anchor length, were calibrated and equated by different methods using the anchor-item design. Comparisons among the equating results yielded by two IRT methods and Tucker's linear equating were presented. Discussions and suggestions were made for future research and equating practice.

Research Purposes

To better understand the function of IRT equating, and to evaluate the adequacy and the effectiveness of the IRT models, IRT results should be compared with the results of classical equating, against some unbiased criteria. Also, the effects of the characteristics of anchor items should not be overlooked. This study, therefore, has the following purposes:

1. To compare the equating results by various IRT methods.
2. To estimate, evaluate, and compare the equating accuracy of the traditional linear equating and the IRT-based equating.
3. To investigate the effect of the test and item characteristics on both linear and IRT equating. Specifically, the effect of anchor length on equating accuracy would be examined.
4. To inform testing practice about the selection of equating methods, based on the findings of the empirical study and the literature review; and to propose useful suggestions on research designs for future studies.

Research Questions and Limitation

The research questions of this study were shaped by both personal interest and the particular context of the test data under study. Because the test scores had already been collected from non-equivalent groups with the anchor-item design, equating based on anchor items was the only choice. The availability, cost, and capacity of computer packages for IRT calibration also set limits on the design of this study.

Taking into account the research purposes and limitation, the following research questions were raised:

1. Was there difference between the equating results of different IRT methods, namely, the two-stage and the fixed-b equating?
2. To what extent that the IRT equating results agreed with the linear equating results?
3. How adequate were the linear equating and the IRT equating, in terms of the accuracy of equating?
4. What constituted a better criterion, for the particular test data used in the study, for evaluating the equating accuracy?
5. Did the equating results depend on the length of the anchor? Specifically, did the equating accuracy improve with the increase in the number of anchor items? What was the most efficient anchor, for the purpose of equating, that had minimum but optimal number of item?
6. Was it appropriate to include the guessing parameter for a minimum competence test, where the score distribution was negatively skewed?

Literature Review

Underlying assumptions and necessary conditions of equating, common equating designs and models, as well as applications of equating in practice are reviewed in the following section.

Conditions of Equivalency

If test Y is to be equated to test X, no matter what equating procedure is chosen, the following conditions must be satisfied to conclude that the scores on test X and test Y are equivalent

(Lord, 1980; Angoff, 1984; Petersen, Kolen, & Hoover, 1989; Dorans, 1990):

1. Both tests measure the same construct.

2. The equating achieves equity. That is, for individuals of identical proficiency, the conditional frequency distributions of scores on the two tests are equal.

3. The equating transformation is symmetric. That is, the equating of Y to X is the inverse of the equating of X to Y.

4. The equating transformation is invariant across sub-populations of the population on which it is derived.

In addition to the above conditions, unidimensionality is also a necessary requirement for the applications of IRT equating. The conditions of equivalency are further explained below.

Same Construct

The requirement of the same construct is a matter of test construction. It can be achieved by carefully selecting items that measure the same construct during the test construction process. When it is desired to compare tests measuring different constructs, equating is achievable but meaningless, because it is simply a problem of regression or prediction. Since equating is a matter of transforming scores for the purpose of comparison, it makes no sense for the forms of a test to measure different constructs.

Equity

The condition of equity requires that individuals of the same proficiency obtain the same scores, no matter what tests are taken. That is, at every ability level, the conditional frequency distribution on one form is the same as that of another form. For equivalent scores, the corresponding percentile ranks in any given group should be equal. The proficiency of individuals taking two different tests are usually estimated via their performance on the common items or an anchor test.

Symmetry

The score transformation should be invertible to achieve symmetry. To say that the scores on test X and test Y are equivalent, regardless of equating from X to Y or Y to X, the same score on one test should correspond to a given score on another.

Population Invariance

It is desired that the equating results be independent of the unique characteristics of the examinee samples used in equating process. No matter which groups of examinees are used, the equating results should not change with the characteristics of the particular examinee groups, except for the underlying construct that the test is measuring. For all the examinees with the same ability, IRT equating is expected to assign them the same estimated ability level. The condition of population invariance is one of the ultimate goals of test equating.

The condition of invariance can be assessed by examining the relationship of equivalence across sub-groups. If population invariance is not obtained, the tests or test forms may not

measure the same construct. As a result, the test construction procedures and test items should be scrutinized.

Unidimensionality for IRT Equating

Although unidimensionality is not explicitly recognized as a condition of equating, it is, however, an underlying assumption for equating based on item response theory. As a result, IRT equating is more restrictive by requiring unidimensional test items.

Overall Equating Guidelines

There is no absolutely superior criterion for the selection of the equating design or method. Judgments and decisions based on equating expertise and experience are needed through out the process. Factors such as feasibility, cost, and the unique testing context should be considered. Because decisions and judgments were arbitrary, Brennan and Kolen (1987) provided a set of guidelines for satisfactory equating.

For test structure, they argued that the test content and statistical specifications for tests being equated ought to be defined precisely and be stable over time. Item statistics should be obtained from pretest or previous use of the test, in the process of test construction. The test should be reasonably long, with at least 35 items, and the scoring keys should be consistent. The stems for common items, alternatives, and stimulus materials should be identical for the form to be equated from and to.

Brennan and Kolen further listed ideal situations for equating as follows: (a) Two sets of common items embedded in the full-length test were desired; (b) The anchors should be at least

1/5 of the total test in length and mirror the total test in content specification and statistical characteristics; (c) At least one link form was administered no earlier than one year in the past, and at least one link form was administered in the same month as the form to be equated; and (d) Each common item was in approximately the same position in the two forms.

They also indicated that the characteristics of examinee groups should be stable over time. The sizes of the groups should be relatively large, roughly speaking, larger than 400. The curriculum, training materials, and field of study should be stable. The test items should be administered and secured under standardized conditions.

Criteria for Selecting Equating Methods

How shall we select or tailor an equating method to our particular needs? For instance, if guessing is explicitly encouraged for test takers and the effect cannot be neglected, a fair equating should account for the factor of guessing. Suppose the equivalent scores are expected to be highly accurate to legitimate its use in certifying professionals, it is critical to select an equating method that functions better for that particular test.

There are three aspects, in general, to consider in the selection of equating method. They are as follows: (1) Are the underlying assumptions tenable? (2) Is the procedure practical? and (3) How good is the equating result? (Crocker & Algina, 1986)

Tenability of Model Assumptions

The premise of a model application is that all the underlying assumptions hold. Linear equating assumes that the tests being equated have identical shapes for the score distributions. It is therefore appropriate to be used when score distributions differ in the means and/or standard deviations only. The derived equivalents will have same percentile ranks due to the assumption.

Equipercentile equating requires fewer assumptions and does not assume the same shapes of score distributions. Thus it is applicable even when the percentile ranks of the two examinee groups are different. The equating procedure determines which scores of different tests have the same percentile rank, instead of assuming the similarity in the ranks. Compared to linear equating, equipercentile method associates with larger errors and the procedure is more complicated.

Both equipercentile equating and linear equating assume that the tests being equated measure the same trait with equal reliability. If the assumption does not hold, the two equating methods may lead to erroneous results. Nevertheless, given two tests of different difficulty, the assumption of equal reliability usually does not hold. In addition, the results of the two methods depend on the particular test items used and fail to meet the condition of equity for equating equivalency. Hambleton and Swaminathan (1990) further indicated that the methods did not meet the requirement for group invariance. Unlike the other methods, IRT equating does not have the above drawbacks and may be a better alternative.

Feasibility of Design and Method

Current equating designs differ in terms of the need of randomly selected groups, the administration of tests, and the use of anchors. If random assignment is employed to form equating groups or the groups take both tests in randomly assigned orders, classical equating will be adequate. Otherwise, IRT-based methods are more appropriate.

Random assignment may save time or money for equating, but it is not always practical or feasible because tests are usually administered to convenient intact groups of examinees. One solution is the use of anchor design, using either anchor items embedded in both tests (the internal anchor) or administering a third test to both examinee groups (the external anchor). Without random assignment, the anchor score distributions for different sub-populations may be markedly different. Thus the assumption of equity is unlikely to hold (Crocker & Algina, 1986). If either linear or equipercentile method is used, the result is unlikely to be accurate. Methods based on latent trait theory are more costly but tend to be more adequate when there is no random assignment. Therefore, they are used most often in such design.

Equating Accuracy

A major concern for test equating is to what extent the equated scores are equivalent. The accuracy of equating depends on the conditions of equivalency (same construct, equity, symmetry, and group invariance).

Since the true score cannot be known and can only be estimated from the observed score, perfect equivalency can never

be determined. Consequently, there is no absolute criterion for equating accuracy, so the degree of accuracy is often studied by comparing the equating result against arbitrarily sound criteria. Equating accuracy is therefore an estimate that depends on the nature of the arbitrary criteria used.

It is unreasonable to compare all equating results against one single criterion because different equating vary in the characteristics of the particular test forms, model assumptions, and equating procedures. There are several ways for selecting criterion for equating accuracy. The equivalent scores derived from conventional equating methods that function well or have been applied for some time can be used as the criterion. The test itself may also form a criterion, particularly in the study of scale drift where a test is equated to itself. Based on empirical studies, IRT-based equating results seem to be more accurate. However, whether the criteria of accuracy is unbiased should remain under scrutiny. Commonly used criteria for equating accuracy is discussed later with other issues on equating accuracy.

Tucker's Linear Equating

Linear equating is appealing for its requirement of a simple linear transformation of raw to scaled scores. Among many the linear equating methods, the popular Tucker's linear equating is employed in this study to compare to the results from IRT equating.

Kolen and Brennan (1987) formulated Tucker's linear equating by emphasizing the notion of a synthetic population, a combination

of the proportionally weighted populations of examinees taking different test forms. Given the total score (X) on one test form, the total score (Y) on another test form, and the total score (V) on the set of anchor items, Tucker's equating makes strong statistical assumptions as follows:

1. The linear regression function (slope and intercept) for the regression of X on V is the same for the two populations. A similar assumption is made for Y and V.
2. The variance of X given V is the same for the two populations. Similarly, the variance of Y given V is the same for the two populations.

With the assumptions on the variance and regression functions in relation to the two populations, Tucker's linear equating is more accurate when groups are similar. As a consequence, the linearly transformed scores on one form have the same mean and standard deviation as scores on another test form.

Though equally reliable test forms are often needed for Tucker's equating, Kolen and Brennan (1987) argued that if the tests were designed to be as similar as possible in content and statistical characteristics and to be equal in length, small differences in reliability between the test forms were not likely to have negative influence on the equating results.

Advantages of IRT Equating

Traditional equating method can yield good results if the test forms are sufficiently parallel (Lord, 1980). However, when the tests to be equated differ in difficulties, IRT methods are considered better than classical linear methods. Major advantages of IRT equating are summarized as follows:

Curvilinear Equating

IRT methods are capable of modeling either a linear or curvilinear relationship between raw scores on two editions of a test. It makes no assumption of equal reliability or identical observed score distributions (Cook & Eignor, 1983; Kolen, 1981). The result of IRT equating often agrees with linear equating to a surprising degree. One possible explanation is that the test construction has already produced considerably similar tests (Berk, 1982).

Item-Free & Person-Independent Measures

The most prominent advantage of the IRT method is the possibility of getting "item-free" estimates for persons and "person-free" item characteristics (Lord, 1977). Ideally, examinees of same ability will get the same ability score, no matter which items are taken. The IRT method can automatically equate different tests or tests forms while calibrating the test items on the same scale.

In addition, IRT models provide estimated error of measurement for ability estimation at each ability level, while classical equating methods only yield a single standard error of measurement for all examinees. Green, Yen, and Burket (1989) suggested that the IRT method would yield equivalent ability estimates for item sets differing in difficulty and/or discrimination, although the equivalent estimates might associate with different standard errors of measurement.

Practical Appeals

IRT equating also have the following practical advantages:

(1) It provides better equating at the upper end of the score scale, where important decisions are often made.

(2) It improves the flexibility in choosing editions of a test, once the editions are placed on the same scale.

(3) If re-equating is necessary, usually after adding or dropping certain items, it is easier to obtain the true score estimates.

(4) It enables pre-equating, which derives the relationship between the test editions before they are administered operationally, when pretest data are available (Cook & Eignor, 1983).

(5) For test forms across years that differ somewhat in content and length, bias or scale drift in equating chains of circular-equating paradigm may be reduced, and the stability of the scales near the extreme values will increase. (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

Based on the above arguments, IRT-based equating methods seem to be superior than the classical methods. However, the relative efficacy of IRT application remains uncertain for the lack of an absolute criterion for equating adequacy. Many questions are await to be answered. For example, how shall the efficacy of equating be estimated? Is the equating result sensitive to various item characteristic curves, given its arbitrary nature in the origin and the scale? If different iterative procedures were used, would the results agree?

Green, Yen, and Burket (1989) noted that it was not safe to say that the IRT method would yield equivalent ability estimates if the item sets were different in content coverage. On the other hand, it was shown that content variations had substantially smaller effects on ability estimates than it had on item parameters (Yen, 1980). The effects of content variations on ability estimates is not clear when the content differs substantially. Therefore, content equivalency should be achieved before equating.

IRT Equating Methods

IRT equating depends heavily on the particular calibration process. The following section provides an overview of IRT equating methods and the role that calibration plays in equating. Generally, IRT equating involves four steps (Hambleton and Swaminathan, 1990):

(1) Choose appropriate equating design with respect to the nature of the test and the group of examinees.

(2) Determine appropriate item response model and assessing model-data fit by broadly gathering goodness-of-fit measures, including statistical tests of significance and checks of model assumptions.

(3) Establish a common metric for ability and item parameters by determining the equating constants for relating either ability parameters or item parameters.

(4) Make decisions on the scale to report the test scores; either ability scores, estimated true scores, or observed scores may be used.

The commonly used IRT methods are introduced below.

Two-stage Method

The two-stage IRT method applies to the anchor test or anchor item design. First, both test forms are calibrated separately. Then the forms are equated using the information based on the anchor items. The following steps are generally involved:

(1) Assume that all items, including the anchor items, measure the same latent trait. Estimate the item difficulties (b's) for the items on the two forms (Book-A and Book-B) of the test separately, using all the items including the anchor items. In the process of calibration, fix the average ability score to zero to set the scale for the item estimates.

(2) Using the anchor items only, compute respectively the average item difficulties for Book-A and Book-B.

(3) Compute the difference of the mean anchor item difficulties by subtracting the mean anchor item difficulty of Book-B from that of Book-A.

(4) For subjects taking Book-A and Book-B, estimate their abilities separately. These estimated ability scores are expressed on two different scales. Since the two scales are expressions of the same latent trait, the scales must be related by the equation $\Theta_1 = \Theta_2 + m$, where Θ_1 is the scale for the group of examinees taking Book-A, Θ_2 is the group of examinees taking Book-B, and m is the mean difference of item difficulty.

(5) Add the mean difference of item difficulty (m) to the ability estimates of subjects taking Book-B. The rationale is that since the numeric value of item difficulty is a value on the

ability scale, if $\Theta_1 = \Theta_2 + m$, then $b_1 = b_2 + m$. By adding m to the ability score on the scale for the group taking Book-B, the ability scores of the examinees taking Book-B are transformed to the scale of Book-A (Crocker & Algina, 1986; Hambleton & Swaminathan, 1990).

Fixed-b Method

The fixed-b method sequentially calibrates the test items by the following steps:

(1) Estimate b's and other item parameters for the Book-A items;

(2) Calibrate Book-B items by fixing b's of the anchor items at the values obtained from the previous step;

(3) Book-B scale is then fixed on to the scale of Book-A (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

True Score Equating

The values on the Θ scale may be transformed to their corresponding true score values when reporting Θ is not preferred. The true score of an examinee with ability Θ on a test is the sum of the conditional probabilities of correct responses across the item characteristic curves. It is defined as follows (Lord, 1980; Crocker and Algina, 1986):

$$\text{True score } (\xi) = \sum_{1}^{n} P_i(\Theta);$$

where Θ is the ability and n is the number of items.

Theoretically, it is possible to equate the true scores on two tests. Suppose the ability level of an examinee on test X is Θ_x and ξ_x is the corresponding true score, and the ability level of

the same examinee on test Y is Θ_y and the true score is ξ_y. Then,

$$\xi_x = \Sigma\ P_i(\Theta_x), \text{ and}$$
$$\xi_y = \Sigma\ P_j(\Theta_y) \equiv \Sigma\ P_j\ (\alpha\ \Theta_x + ß\ );$$

where $\Theta_y = \alpha\,\Theta_x + ß$ depicts the linear relationship between Θ_y and Θ_x.

Therefore, for a given value Θ_x, the pair of true scores (ξ_x, ξ_y) on the tests X and Y is determined (Hambleton & Swaminathan, 1990).

To emphasize the IRT model and data fit, Hambleton, Swaminathan, and Rogers (1991) substituted ξ with τ and rewrote the equation as follows: $\tau|\Theta = \sum_{1}^{n} P_i(\Theta)$. The τ is called test characteristic curve (TCC) for it is the sum of the item characteristic curves. Since each $P_i(\Theta)$ is an increasing function of Θ, τ and Θ are monotonically related. The larger the Θ, the larger the corresponding τ. The range of τ is between 0 and n, and it is on the same scale as the number-right scale. Lord (1980) indicated that true score ξ and ability Θ are the same thing expressed on different scales of measurement. The difference is that the scale for ξ depends on the number of items on the test; the scale for Θ is independent of the number of items on the test. Therefore, Θ is more useful than ξ for comparison purposes.

Guessing occurs frequently for multiple choice tests. The probability of guessing an item right depends on the number of the alternative options for that item. Taking into account the number of alternatives for each item, the true formula score can be defined with the following equation and the test characteristic curves for Book-A and Book-B can be formed:

$$\text{True score } (\xi) = \sum_{1}^{n} \{[(k_i+1)/k_i]\ P_i(\Theta) - 1/k_i\};$$

where n is the number of test items, and (k_i+1) is the number of choices for item i (Petersen, Cook, & Stocking, 1983).

Concurrent Calibration Method

Using LOGIST, an IRT calibration computer program on mainframes, the item and ability parameters can be estimated simultaneously in the following manner:

(1) Treat examinees taking Book-A and Book-B as one sample. Treat data as if all the examinees have taken a test consisting all the items from both Book-A and Book-B.

(2) Since the examinees taking Book-A do not respond to the items on Book-B, code the scores for Book-B items as "not reached" for the examinees taking Book-A. Treat the scores for Book-A items for the examinees taking Book-B similarly.

(3) Calibrate in a single LOGIST run for the ability parameters for all the examinees and the item parameters for all the items. The ability estimates for the examinees taking either Book-A or Book-B are automatically put on the same scale. No further step is needed (Hambleton & Swaminathan, 1990).

Conceptually, the concurrent calibration method is expected to yield more stable equating results because it does not make any assumptions about the relationship between the item parameter scales for separate calibration runs (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

Characteristic Curve Transformation (Formula Method)

The steps of the characteristic curve transformation method

are as follows:

(1) Book-A and Book-B are analyzed separately to obtain two sets of item parameters.

(2) For each form of the test, calculate the mean and standard deviation of the b's for the anchor items.

(3) To put the parameters of Book-B on the scale of Book-A, linearly transform the Book-B item parameters using the following formulas (Stocking & Lord, 1984; Hambleton & Swaminathan, 1990, p.205, 222):

$$b_y = \alpha b_x + ß,$$
$$a_y = a_x / \alpha;$$

where $\alpha = S_y/S_x$ and $ß = \bar{y} - \alpha\bar{x}$, $\bar{y}$ and $\bar{x}$ are means, and S_y and S_x are standard deviation of b-values for the common items.

The basis for the linear transformation is that, in anchor test design, the difficulty and discrimination parameters for the common items are linearly related between the two tests, assuming item and people invariance. (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

IRT Pre-equating

Item pre-equating design establishes equating conversions between a new test edition and a previous one prior to administering the new edition. It depends on adequate pre-testing of a pool of items where the new edition is assembled. The pre-testing is conducted while the editions already equated are operationally administered. Item statistics obtained are used to equate scores on the new edition to desired scale (Cook & Eignor, 1983). To better pre-equate, items must behave similarly in

pretest and operational contexts, especially for item position and influential context effects. It was found that pre-equating method associated with substantially larger bias and errors (Kolen & Harris, 1990).

The Use of IRT Equating Coefficients

The IRT equating transforms the values of the item and examinee parameters on one metric into those of the other or a base metric. Therefore, the slope and intercept coefficients of the appropriate linear transformation of the parameter estimates must be obtained. For examinee parameters, the basic transformation equation is as follows (Baker & Al-Karni, 1991):

$$\Theta^{*} = A\Theta + K$$

where A is the slope, K is the intercept, Θ is the examinee's ability parameter in the metric to be equated, and Θ^{*} is the converted Θ on the target metric.

For the item parameters, the transformation can be done as follows:

$$a^{*} = a/A$$
$$b^{*} = Ab+K$$

Sampling Effects on Equating Results

Sample invariance is a desirable property of equating method. Ideally, we would like to obtain equating independent of ability level for any sub-population. Although it was argued that equating results were consistent across sub-populations of different ability in general, Lawrence and Dorans (1990) noted that the results relied on the examinee samples of approximately equal ability. They suggested that population independence should be investigated under circumstances that the examinee samples

differed in ability. The suggestion is relevant because, in reality, examinee groups are more than likely to have different ability.

Effect of Ability Difference

Ability difference between examinee samples may have serious impacts on equating results (Cook, Eignor, & Schmitt, 1988). Theoretically, the closer the groups in the ability being measured, the more accurate the equating will be. That is, the ability estimate will be closer to the true ability score, and examinees with the same ability will get the same score.

To overcome the threat of ability discrepancy between sample groups, particular sampling strategies are employed from time to time to draw samples of similar ability. It is intriguing to learn what combination of sampling procedure and equating method works best, since sampling effect may vary with equating methods. Literature generally recommended the use of representative sample instead of matched sample.

Representative vs. Matched Sampling

Dorans, Livingston, Wright, and Lawrence (1990) found that "matched sampling", stratifying samples on the anchor test, was useful in abridging the disagreement among equating methods. When populations differed in ability, however, matched sampling was not only complicated but also yielded little improvement for equating. Schmitt, Cook, Dorans, and Eignor (1990) also had a similar finding after investigating the sensitivity of equating results to different sampling strategies. Eignor, Stocking, and Cook (1990) recommended that the "matched" samples should not be used with 3PL

IRT equating procedures, because it introduced more errors in the estimates of item difficulty than representative samples.

Equating Tests with Skewed Distributions

Practitioners or researchers are usually interested in equating large scale achievement tests that have approximately symmetrical and bell-shaped score distributions. However, we are often required to equate tests that have highly skewed score distributions such as minimum-competency tests or licensure exams with high passing standards. In addition, especially for licensure and certification programs, test forms are often equated with special interest on a particular cut-off score or range of scores to inform decision making. In this case, the equating procedure itself is not relevant to the procedure of determining the cut-off criterion. To maximize the precision of the decision, it is reasonable to direct more attention to equating in the cutting score region, even at the expense of poorer equating at other scores (Brennan & Kolen, 1987).

Hills, Subhiyah, and Hirsch (1988) equated the scores of a minimum-competency test, the Florida Statewide Student Assessment test, to the scores of an early version that was administered two years before. The test items were from the same content domain, item difficulty were similar, and the examinees were essentially from the same population. It was found that almost all the five equating methods used in the study yielded similar results. Hills, Subhiyah, and Hirsch concluded that IRT equating methods could be used for minimum-competency tests of extremely skewed distributions and yielded reasonable results.

Assessing Equating Adequacy

Equating effect can be evaluated in terms of accuracy, sample invariance, or scale stability. Sample invariance, a desirable result for equating tests, can be assessed by examining the similarity of the equating results obtained from diverse groups, which can be different in ability, socio-economic status, race, or other characteristics. As sample invariance has been discussed previously, in this section, only equating accuracy and scale stability will be reviewed.

Measures of Equating Accuracy

The purpose of equating is to obtain comparable scores that well estimate the underling true score, therefore, a relevant question is: How good are the true score estimates and to what extent the equated scores are comparable? It was argued that IRT methods was superior for its capacity to equate both parallel and non-parallel tests or forms (Kolen, 1981). Green, Yen, and Burket (1989) found that IRT-based procedures were effective for both inter-level and inter-form equating. Unfortunately, the findings are tentative because the accuracy measure can only be determined with an arbitrary criterion.

This section focuses on how a criterion equating is selected in practice and how the evaluation on equating accuracy can be done. Brief description of the commonly used accuracy measures for equating is included.

Selection of criterion equating.

If certain conventional equating methods are known to function well or have been in existence for some time, the results

of the conventional methods may be used as a criterion against which the IRT equating is evaluated. For example, in a comparative study, Livingston, Dorans, and Wright (1990) made an assumption that the true equating relationship was the equipercentile relationship in the target population because the true scores could be precisely estimated. Yen (1985) also suggested the use of the equipercentile equating as the target equating for its equal accuracy to the IRT methods.

Equating to self.

In other situations, the test itself may form a criterion accuracy. For example, a test is equated to itself in a typical design for studying scale drift. Skaggs and Lissitz (1986) concluded that the best situation for research purposes occurred when a test could be equated with itself through intervening forms. Yet one must be cautious when interpreting the results, because time has elapsed between test administrations and equating error could be confounded with other types of measurement errors.

Estimated RMSD and BIAS.

A common overall accuracy measure for equating is root-mean-squared deviation (RMSD), also called root-mean-squared error of equating (RMSE). It is based on the residual of the equated scores from the criterion accuracy measure for the full sub-population. The criterion accuracy measure is obtained from a criterion equating or the corresponding raw score (Klein & Jarjoura; Livingston, Dorans, and Wright 1990, 1985). Suppose Form-B of a test is equated to Form-A, then

$$RMSD = \{ [\sum_{y=1}^{y} n_y (\hat{x}_y - x_y)^2] / \Sigma n_y \}^{1/2} ,$$

where n_y is the number of examinees with raw score y on Form-B; x_y is the corresponding exact scaled score on Form-A as determined by the criterion equating; and $\hat{x}_y$ is the corresponding exact scaled score on Form-A as determined by the equating that is to be evaluated. The summation is over the raw-score levels on Form-B.

The formula can be rewritten as: $RMSD = \{ [\Sigma(\hat{x}_i - x_i)^2] / n \}^{1/2}$.

Klein and Jarjoura (1985) estimated the mean equating error, the bias that contributes to the RMSD, with the following formula:

$$BIAS = \bar{X} - \bar{X}';$$

where $\bar{X}$ is the mean of the criterion scores and $\bar{X}'$ is the mean of the equivalents.

Livingston, Dorans, and Wright (1990) computed a variation of the bias statistic to diagnose for a large RMSD. The statistic is a weighted mean difference for the new-form population as shown below:

$$Bias = \Sigma [n_y (\hat{x}_y - x_y)] / \Sigma n_y$$

It is noted that the bias measure is not good at evaluating an equating unless all the equated scores were too high or too low. However, it describes the tendency to produce equated scores that were systematically too high or too low.

Marco, Petersen, and Stewart (1983) investigated the adequacy of a variety of curvilinear score equating models on verbal portion of the SAT. The criterion equating was the test score itself when a test was equated to itself. In the cases where

tests were equated to a different test, two criteria were established: a. equipercentile equating of observed scores (the direct equipercentile criterion); and b. equipercentile equating of estimated true scores derived from the 3PL model (the IRT equipercentile criterion). Because the IRT equipercentile criterion might be biased in favor of the IRT equating methods, the direct equipercentile criterion was adopted.

Two discrepancy indices were used to evaluate the effectiveness of the models in the study of Marco, Petersen, and Stewart (1983): the standardized weighted mean square difference (or the total error) that gave the greatest weight to those values most likely to occur, and squared bias.

Marco, Petersen, and Stewart (1983) warned for the tentativeness of equating results based on arbitrary criterion of equating accuracy. However, they found that if the anchor test mirrored the content and the difficulty level of the total test, the sample differences had relatively small and unsystematic effects on the quality of the equating results. They also found that internal anchor, the common items embedded in the tests being equated, resulted in less total error than external anchor, which was a third test given to both of the examinee groups. They attributed it to the possible difference between the external anchor and the total test.

Scale Stability

Though stability is different from accuracy, it is sometimes used to compare the adequacy of different equating methods. The

BEST COPY AVAILABLE

research findings of scale stability are briefly summarized below, since this study does not focus on scale stability.

Kolen (1981) investigated the stability of equating results across stratified random samples via a cross-validation design using nine equating methods, including classical method and IRT models. The cross-validation criterion used was a mean-square-difference index, of which a smaller value reflected greater consistency. He also computed Friedman statistic to conduct an overall significance test for differences among various equating methods. Kolen concluded that the one parameter IRT models were inadequate possibly because the prevalence of guessing by the examinees. In addition, the 3PL IRT model seemed adequate, so was the equipercentile method. However, Kolen noted that these conclusions were tentative because the sampling distribution of the cross-validation statistic was unknown and the consistency among the methods was only a relative measure of stability. The complex interaction between item content, difficulty level, and the equating model may make the results of the cross-validation statistic even harder to interpret (Skaggs and Lissitz, 1986).

If the result of directly equating the new edition of a test to its old edition is not the same as the result of equating the new edition to the old edition through intervening edition(s), there is scale drift due to equating method (model fit problems) and/or sampling errors. Scale drift often indicates the inadequacy of equating methods. Brennan and Kolen (1987) indicated that equating procedures estimating only one or two moments tended to be more appealing than the procedures estimating many moments. In addition, if arbitrary test forms were chosen to

be equated to itself using the circular equating paradigm, the equating results might be different. They attributed the difference to the circular equating paradigm, but not the merits of different equating methods.

The weighted mean square difference was used by Petersen, Cook, and Stocking (1983) as a summary index for evaluating the effectiveness of the various equating models. The measure gave greater weights to those values that were most likely to occur and represented larger discrepancies, as shown in the following formula:

$$\Sigma_j f_j d_j^2 / n = \Sigma_j f_j (d_j - \bar{d})^2 / n + \bar{d}^2 ;$$

(Total error) (Variance of Difference) (Squared Bias)

where

(1) $d_j = t'_j - t_j$; t'_j is the estimated scaled score for raw score x_j, and t_j is the initial or criterion scaled score for x_j;
(2) f_j is the frequency of x_j;
(3) $n = \Sigma_j f_j$; and
(4) $\bar{d} = \Sigma_j f_j d_j / n$.

The summation is over that range of x where extrapolation is unnecessary (Petersen, Cook, and Stocking, 1983).

Overall, it was found that the IRT conversions had less discrepancy from the initial scale than the other equating methods. Methods based on the three-parameter logistic IRT model resulted in greater stability of equating when tests differed somewhat in content and length (Petersen, Cook, & Stocking, 1983).

Assumption of Dimensionality

The robustness of IRT models to violation of assumptions is a major concern of the IRT test equating application. Among them,

multidimensionality affects the fit of an IRT model most. It is especially true for achievement tests where several different types of content are tested.

Effect of Dimensionality

In practice, test scores are most meaningful when all the items depend on a single trait. If the IRT assumption of unidimensionality holds, local independence should be observed; that is, for fixed Θ, the item characteristic functions for any pair of items i and j are independent (Lord, 1982). If the probability for a given response to the given items i and j are not independent for fixed Θ, the responses to items i and j depend on some trait other than the Θ shared by the two items. Then there is a violation of unidimensionality.

Robustness of Unidimensionality Assumption

The study of Dorans and Kingston (1985) showed that violation of unidimensionality might have an impact on equating, but the effect might not be substantial. The influence may depend on how the violation of unidimensionality is formulated. It was found that dimensionality influenced the magnitude of item discrimination parameter estimates, which caused an asymmetry of equating. However, given the similar equating obtained from tests with varying dimensionality, IRT equating might be sufficiently robust to the dimensionality violation. Dorans and Kingston suggested that it was reasonable to think that an overall ability, the total verbal ability in their case, could be thought of as a weighted composite of the separate component abilities (the verbal and reading comprehension abilities).

Reckase, Ackerman, and Carlson (1988) also advocated that the unidimensionality assumption was robust by arguing that unidimensionality only required items of a test to measure the same composite of abilities, rather than a single ability. Yen's suggestion is to hypothesize that the unidimensional model chooses a combination of underlying traits as its unidimensional trait (Yen, 1984). If a test involve independent traits that influence only a few items, the traits can be ignored in the definition of the unidimensional three-parameter trait.

Moreover, Dorans (1990) argued that the tests being equated did not have to be composed of unidimensional items, although they had to measure the same construct. The tests should contain the same content mix of items, and sets of items could be selected to meet the unidimensionality assumption for most IRT models, even when more than one ability was required to give a correct response. Reckase, Ackerman, and Carlson (1988) demonstrated theoretically and empirically that sets of items measuring the same weighted composite of abilities could be selected to meet the unidimensionality assumption for most IRT models.

Characteristics of Anchor items

It is agreed that the characteristics of anchor items are influential to equating results. Consequently, it is crucial to adequately select anchor items. Both content representativeness and anchor length are important characteristics of the test anchor and deserve to be studied. However, because of the focus and page limit of this paper, the issue of content representativeness would

only be skimmed in the following paragraph, and only the issue of anchor length would be reviewed extensively.

Content Representativeness

Whether the anchor items are representative to the overall items of the tests being equated, in terms of content and statistical properties (Cook & Petersen, 1987) is especially important when groups vary in ability. Budescu (1985) pointed out that the magnitude of the correlation between the anchor test and the unique components of each test form was the single most important determinant of the efficiency of the equating process. Brennan and Kolen (1987) further indicated that any substantial content changes entailed a re-scaling and re-norming of the test with a new "origin" form to which subsequent forms were equated.

Length of Anchor

If efficiency is considered, it is natural to expect an anchor to be shorter in length but yield better equating result. The reason not to have too many anchor items is to preserve the flexibility in selecting non-anchor test items to reflect the content domain being tested.

Although there is no absolute standard for appropriate length of an anchor, a rule of thumb is given by Angoff (1984) as follows: At least 20 items or 20% of the total number of items in a test, whichever is larger. Several studies suggested, however, that as few as five or six carefully chosen items could perform as satisfactory anchors in IRT equating when the item parameters of both tests were estimated in a single analysis using IRT concurrent method (Raju, Edwards, & Osberg, 1983; Wingersky &

Lord, 1984; Raju, Bode, Larsen, & Steinhaus, 1988; Hills, Subhiyah, & Hirsch, 1988).

It is impossible to offer universal guidelines for selecting the length of the anchor. For its specific purposes, each testing program needs to take into account the time, cost, and context constraints as well as the particular index of efficiency when determine the length of the anchor (Budescu, 1985).

Hills, Subhiyah, and Hirsch (1988) studied the effect of anchor test length and found that five randomly chosen anchor items of a mathematics test was not sufficient to produce satisfactory equating result. However, an anchor of ten items were found satisfactorily sufficient when IRT concurrent method was used.

Precaution: Limits of Equating

Test equating cannot solve problems originated in rough and improper test construction. It should be used, conversely, to overcome the insufficiency of a fair test construction that fails to yield parallel forms.

Both classical and IRT equating are primarily designed for minor differences in difficulty between test forms. Cook and Eignor (1991) indicated that no equating method could satisfactorily equate tests that were markedly different in difficulty, reliability or test content. From this perspective, the practicality of vertical equating, which transforms scores across levels of achievement (usually school grades) onto a single scale, is in question. Theoretically and operationally, vertical equating is much more difficult to accomplish than horizontal

equating. Skaggs and Lissitz (1988) suggested that multidimensionality might account for a lack of test equating invariance of vertical equating.

Equal reliability is usually assumed for equating. Both linear and equipercentile equating require equally reliable tests. Due to floor and ceiling effects, however, tests differing in difficulty are not likely to be equally reliable for all subgroups of examinees, and the relationship between the tests is nonlinear (Skaggs & Lissitz, 1986). It is implied that observed scores on tests of different difficulty cannot be equated. Equating is, in fact, done in its loose sense. From a pragmatic point of view, equating is to arrive at a conversion equation that approximates an ideal equating. Despite its limitation by nature, test equating is still of great use in comparing scores on test forms of minor differences.

The focus of this study is horizontal equating, a permissible and frequently used equating. The test forms being equated were constructed to be parallel.

Description of Data

The test data used in this study were the scores on the two forms, Book-A and Book-B, of a 1993 in-training examination taken by the candidates of a medical specialty. The candidates took the test, while participating in various in-training programs located at different sites (usually in hospitals), to prepare for the board certification examination. No absolute score was used to determine pass or fail. The passing standard was 75% of the total

test items being correctly answered.

To become board-certified, the candidates were strongly motivated to participate in the in-training programs for the preparation of the certification exams. Since the in-training test provided candidates valuable opportunities to get familiar with the formal certification exams, it was assumed that the candidates had taken the test as serious as when the formal exams ware taken.

Test Content and Format

The test forms were comprised of five-alternative multiple-choice items, and the content of all the items were emergency-medicine-related. The item responses were all scored as right or wrong (coded as 1 or 0). Book-A had 203 items, of which 58 items were unique to Book-A. There were 52 unique items in Book-B, and the total number of items was 197. There were totally 145 anchor items, and the anchors were identically embedded in both forms in terms of wording and location.

Examinee Groups

A total of 2,242 candidates took the in-training test. After screening the data, a case that had apparently guessed throughout the entire test was deleted from the analysis to secure the validity of scoring. Among the 2,241 subjects, 1,092 took Book-A and the rest of 1,149 took Book-B.

The examinee group taking Book-B scored higher in average on the anchor items, therefore it was likely that this group of examinees had higher ability. Nonetheless Lord (1981) mentioned, the difference in ability level would not influence equating

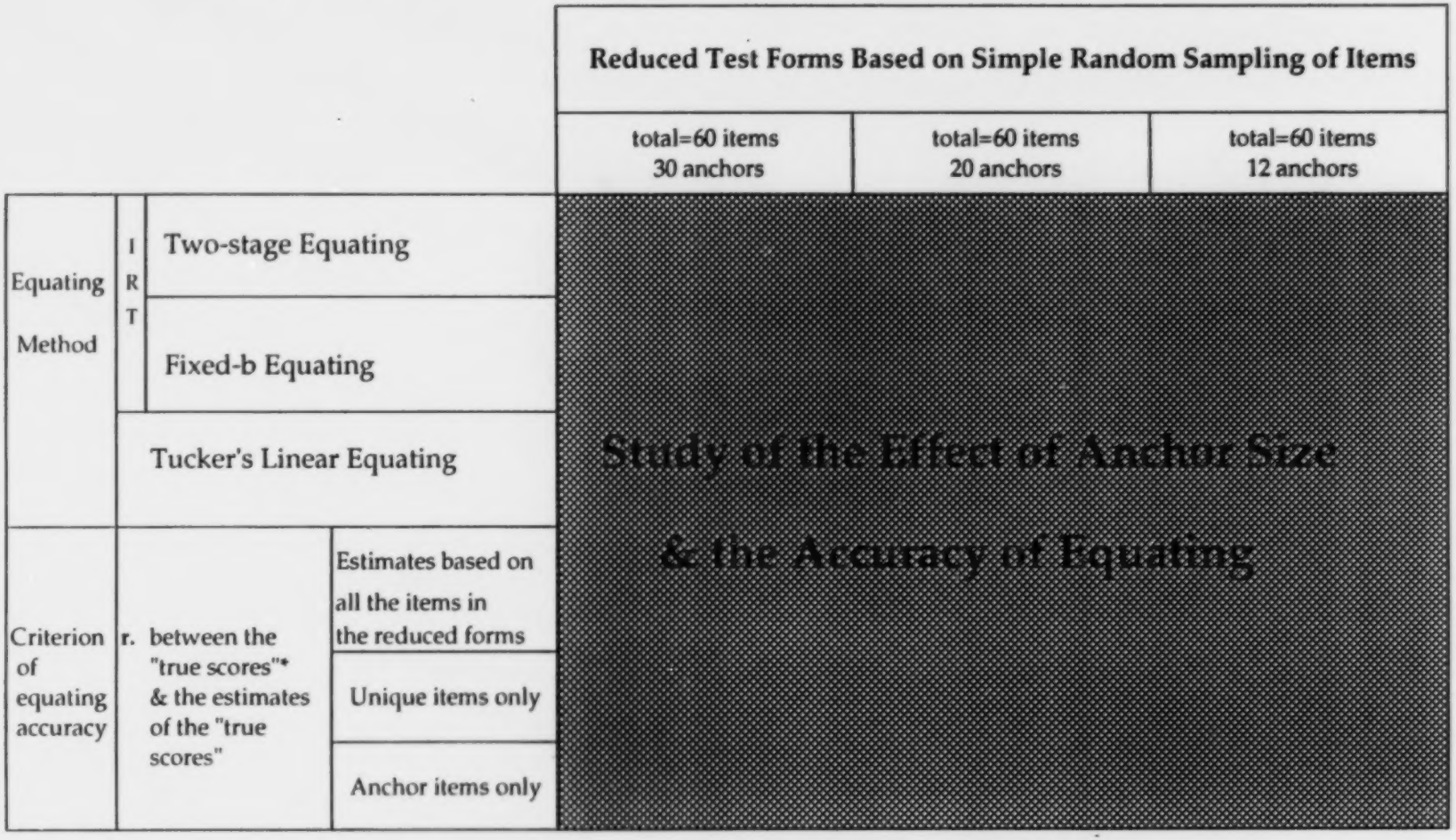
result, given anchor-test design was employed. In addition, the group taking Book-B had a lower mean score on the unreduced full-length test. This implied that the unique items in Book-B had higher difficulty in average.

The test forms generally met the equating requirements that were mentioned earlier in the review of equating guidelines. Specifically, the test was reasonably long and all the items were from one single item pool. The anchor items constituted the major part of the total test. Some of the items were administered in the previous year under the same standardized testing situations. The size of the examinee groups, over 2,200 subjects, were reasonably large. In addition, the scoring key was clear and the test results appeared to be stable, given the preliminary analyses based on the classical test equating.

Research Design

All the equating in this study were based on non-equivalent populations, random sampling of items, and the internal anchor-item design. The following variables, anchor length and equating model, delineated the entire study. The equating results of IRT-based models were compared against the results of linear equating, with a raw-score-based criterion of equating accuracy. The number of anchor items was systematically manipulated to reflect the common suggestions for the anchor length. FIGURE 1 illustrated the basic design of research in this study.

Figure 1: Basic Research Design


Reduced Test Forms Based on Simple Random Sampling of Items
total=60 items 30 anchors
total=60 items 20 anchors
total=60 items 12 anchors
Equating Method
IRT
Two-stage Equating
Fixed-b Equating
Tucker's Linear Equating
Criterion of equating accuracy
r. between the "true scores"* & the estimates of the "true scores"
Estimates based on all the items in the reduced forms
Unique items only
Anchor items only
Study of the Effect of Anchor Size & the Accuracy of Equating

Note: * -- The "true score" in this study is the total raw score for the 145 anchor items from the full test forms.

Anchor Item Design of Equating

The two examinee groups taking different test forms were not formed by random assignment. The test forms were constructed with embedded common items so that the test scores could be made comparable. The internal anchor items were representative to the full-length test, and was embedded in both forms in the same fashion (same wordings and same location).

Random Sampling of Items

The test items fell into 23 sub-areas of a single content domain. Following the research design, subsets of items were randomly drawn from the item pool controlling for the number of anchor items.

The item sampling scheme rendered an opportunity for the study of the anchor length effects. The total number of items drawn for each of the reduced test forms also reflected the common test length in testing practice. The random sampling of items also extended the scope of the study on equating accuracy. As a result, the examinee's performance on the complete set of 145 common items, from the original full-length test, formed a legitimate criterion for equating accuracy.

Specifically, equating results were compared against the total raw score on the complete anchor set. The total raw score was the pseudo "true-score" in the sense that the anchor mirrored the item population and the scores were available for all the examinees. It was indeed the "true score" if the population and the occasion were considered fixed. To evaluate the adequacy of equating results, the estimated true scores of different

procedures were correlated with the pseudo "true score". The Pearson's product moment correlation coefficients were computed as the index of equating accuracy.

The item sampling schemes, the sampling procedures, and the resulted three pairs of reduced test forms were described in APPENDIX 1. In summary, the three item samples that differed in anchor length were drawn from the single item pool. The underlying assumption was that all the items were written for a single content domain. The numbers of common items for the three reduced tests were fixed at 30, 20, and 12 respectively. These figures reflected a considerably long anchor and two anchors of minimal length. The minimum lengths were chosen based on the recommendation that an anchor should have at least 20 items or 20% of the total test items. By using such minimum numbers of anchor items, fewer items from one test to another test were repeated, thus test security could be enhanced (Hills, Subhiyah, & Hirsch, 1988).

It should be noted that, although, from the previous research it was found that five or six carefully chosen anchor items would yield good equating results. For the current research, the study on the effects of a smaller anchor was not feasible. It was because most of the 255 test items were common items and not spread evenly across the 23 content areas.

Equating Methods

In addition to the Tucker's linear equating, the three parameter logistic IRT model was also applied to account for the

guessing factor. Since the test was comprised of multiple-choice items, it was highly likely that the examinees had guessed on some difficult items. Two IRT equating techniques, the two-stage method and the fix-b's method, were used operationally to investigate the effects of IRT calibration on test equating. Although various item samples were created in this study, when comparing equating results of different equating methods, the same set of items was always used for the comparison.

Research Tool

The program used to obtain IRT estimates for item parameter and person ability was the PC version BILOG 3. There were other IRT-based programs, such as LOGIST and ASCAL, that could also be used to calibrate the test items. A comparison between BILOG 3 and LOGIST illustrated why BILOG 3 was used in this study.

BILOG 3 yielded marginal maximum likelihood (MML) estimates, whereas LOGIST simultaneously maximized the joint likelihood function (JML) for the estimates of item and examinee parameters. The JML estimates were likely to become inconsistent when the numbers of examinees or items increased (Mislevy & Stocking, 1989; Baker, 1990). The number of parameters estimated with MML did not increase with the increase in the number of examinees. Consequently, BILOG 3 would yield more consistent results. The marginal maximum likelihood was the probability of making a correct response by an examinee randomly selected from a population with a certain ability distribution. Yen (1987) also found that BILOG always yielded more precise estimates of

individual item parameters. For shorter test with ten items, BILOG excelled LOGIST in estimating item and test characteristic functions. For longer tests with 20 to 40 items, however, the two programs yielded similar estimates.

LOGIST constrained extreme item parameter estimates by imposing specified upper and lower limits. The boundary values affected Θ estimates, even though the effect may be minimal. Based on a Bayesian framework, BILOG implemented prior distributions on all item parameters in the 3PL model. The Θ estimates obtained depend on the characteristics of other items in the test. If the prior information is not appropriate for the data, item parameter estimates would be biased (Baker, 1990). When the tests were longer, samples were larger, and some items were omitted or not reached, LOGIST and BILOG would yield similar item parameter estimates. In this case, BILOG was still more appealing for its cost and statistical properties. If the tests were shorter or the samples are smaller, BILOG had the advantage of yielding more reasonable results (Mislevy & Stocking, 1989).

Unless linear transformation was applied, IRT parameter estimates were not unique when specifying a Θ trait scale metric. The PC-BILOG used the estimated posterior Θ distribution to establish the location and metric for the Θ scale, as a solution to the identification problem. Baker (1990) indicated that the PC BILOG preserved the variability of true distributions that had smaller variances, but it standardized the variability of the true distributions when the variances were larger.

Results and Discussion

The results of the classical item analysis showed that the items had various item difficulty and correlated moderately to the total test score. The results were summarized in APPENDIX 2. Descriptive statistics and correlation coefficients for the common and unique items, mainly for the application of Tucker's linear equating, of each of the test forms were included in APPENDIX 3. Descriptive statistics for IRT calibration for the three pairs of reduced forms were presented in TABLE 2 for both the two-stage and fixed-b methods. The correlation coefficients presented in TABLE 4 described the equating accuracy of different equating methods over various item samples by using the scores on the complete set of 145 anchor items as the accuracy criteria, the "true scores".

Test Homogeneity & Content Representativeness

For the complete test forms and the reduced forms, the anchor and the unique items were significantly correlated. They also correlated significantly with the whole test forms respectively, regardless of the differences in anchor length (see APPENDIX 3). It showed that the anchor reasonably mirrored the test form, in relation to the content. Therefore, the use of the anchor in equating the test forms seemed appropriate.

It should be noted that, however, the magnitude of the correlation coefficient between the anchor and the whole test was inflated by auto-correlation because the anchor was embedded in the test. The artifact of the auto-correlation was evidenced by the decreasing magnitude of the correlation coefficient when the

size of the anchor decreased. The coefficient decreased from .861 to .797 to .729 for Book-A and from .863 to .787 to .696 for Book-B when the number of anchor items decreased from 30 to 20 to 12. Thus the content representativeness of an anchor could not be solely determined by the magnitude of the correlation coefficient. The number of anchor items also played a critical role.

The size of the anchor, therefore, might have substantial impact on test equating. This was discussed later with the IRT equating results.

Preliminary Study on the Raw Scores

Generally, the average raw scores of people taking different test forms did not differ substantially. With the evidence of item homogeneity and content representativeness, the test forms seemed to be pretty parallel. Upon closer inspection, however, it was found that the examinees taking Book-B scored slightly higher on both anchor items and unique items across the three reduced item samples. The average raw scores was divided by the number of items in the form to yield comparable percentage values. The percentages were summarized in TABLE 1.

The figures in TABLE 1 indicated higher ability of the examinees taking Book-B over the three pairs of reduced test forms, perhaps caused by the non-random selection of the examinees. Despite the ability difference between examinee groups, equating results should not be affected because of the anchor-item design.

TABLE 1

Average Item Difficulty ($\bar{p}$): The % of Examinees Getting an Average Item Right

		Simple Random Sampling of Items		
		30 anchor items	20 anchor items	12 anchor items
Anchor Items	Book-A	0.722	0.686	0.667
	Book-B	0.736	0.705	0.691
Unique Items	Book-A	0.654	0.676	0.685
	Book-B	0.670	0.701	0.693

<Note> 1. P_i=the percentage of examinees getting the item "i" right

$\bar{P} = (\sum_{1}^{n} P_i)/n$; n = # of items in the reduced test form

2. Sample size: $N_{Book\text{-}A}$=1092; $N_{Book\text{-}B}$=1149.

Estimation of Item Parameters

The IRT equating in this study were done by incorporating the two-stage and the fixed-b methods with the three parameter logistic IRT model for item calibration. The summary statistics for the estimated item and ability parameters were presented in TABLE 2. The distributions of the item parameter estimates generally showed the items from different test forms had different item difficulty but similar item discrimination. It suggested the item sampling effect on test content and was consistent with the pattern found in the average item difficulty based on raw scores. Since Book-B was equated to Book-A, only the estimated parameters for Book-A were used in computing the estimated true scores later.

Equating the Estimated Ability Scores

The average anchor item difficulty of the form being equated to (Book-A), yielded by the two-stage calibration, was used to adjust for the ability estimates from the form being equated (Book-B). Since the fixed-b method resulted in estimates that were already on the same ability scale, no adjustment was needed.

The equivalent ability scores yielded by the two IRT methods were very similar. The correlation coefficients between the two sets of equivalent ability scores were computed for various reduced forms. The significant high correlation, .999 for all the reduced forms (see APPENDIX 4), showed that the two-stage and the fixed-b equating methods yielded almost identical equating results in terms of individual examinees' standings in the examinee group.

TABLE 2: Summary of IRT Calibration

				Simple Random Sample		
Book-A (n=1092)		Total # of items		60	60	60
		# of anchor items		30	20	12
		Estimated a	mean	0.340	0.365	0.356
			s.d.	0.173	0.179	0.179
		Estimated b	mean	-0.884	-0.544	-0.761
			s.d.	2.239	2.290	2.540
		Estimated c	mean	0.252	0.259	0.249
			s.d.	0.046	0.050	0.050
		Mean anchor item difficulty		-1.340	-0.620	-0.420
		Ability estimates	mean	0.003	0.004	0.004
			s.d.	0.851	0.864	0.860
Book-B (n=1149)		Total # of items		60	60	60
		# of anchor items		30	20	12
	Two-stage Method	Estimated a	mean	0.377	0.421	0.409
			s.d.	0.165	0.193	0.194
		Estimated b	mean	-1.008	-0.628	-0.531
			s.d.	1.891	1.969	2.141
		Estimated c	mean	0.241	0.279	0.277
			s.d.	0.034	0.057	0.057
		Mean anchor item difficulty		-1.450	-0.800	-0.540
		Ability estimates	mean	0.003	0.005	0.005
			s.d.	0.868	0.876	0.871
	Fixed-b Method	Estimated a	mean	0.400	0.429	0.406
			s.d.	0.164	0.173	0.176
		Estimated b	mean	-0.591	-0.073	-0.079
			s.d.	1.951	1.999	2.418
		Estimated c	mean	0.311	0.353	0.341
			s.d.	0.053	0.070	0.068
		Ability estimates	mean	0.059	0.176	0.174
			s.d.	0.880	0.910	0.906

<Note> a = Discrimination Parameter
b = Item Difficulty Parameter
c = Guessing Parameter

Estimation of the True Scores

With item parameter estimates and equivalent ability estimates for both examinee groups, the true score estimates on Book-A for all the 2,241 examinees were obtained by using the following formula (Lord, 1980):

$$\text{Estimated true score } (T) = \sum_{i=1}^{n} P_i(\Theta)$$

$$= \sum_{i=1}^{n} \{c_i+(1-c_i)/\{1+\text{Exp}[-1.7a_i\ (\Theta-b_i)]\}\ \};$$

where Θ is the examinee's ability and n is the number of items.

The correlation between the estimated true scores resulted from the two-stage and the fixed-b equating methods was significantly high, as expected, across all set of test forms.

Tucker's Linear Equating Results

Tucker's linear equating was applied to all the reduced test forms. The results were summarized in TABLE 3.

TABLE 3: Summary of Tucker's Linear Equating Results

For the test forms with 30 anchor items:

$\alpha_A(A|V)=1.5236$ & $\mu_S(A)=41.6151$ & $\sigma^2_S(A)=32.7416$
$\alpha_B(B|V)=1.5925$ & $\mu_S(B)=41.8546$ & $\sigma^2_S(B)=36.0471$

Equating equation: $l(b)=.9530(b-41.8546)+41.6151$

For the test forms with 20 anchor items:

$\alpha_A(A|V)=1.8270$ & $\mu_S(A)=41.0976$ & $\sigma^2_S(A)=34.9185$
$\alpha_B(B|V)=1.8605$ & $\mu_S(B)=41.7714$ & $\sigma^2_S(B)=37.3508$

Equating equation: $l(b)=.9669(b-41.7714)+41.0976$

For the test forms with 12 anchor items:

$\alpha_A(A|V)=2.1554$ & $\mu_S(A)=41.1826$ & $\sigma^2_S(A)=32.6102$
$\alpha_B(B|V)=2.2419$ & $\mu_S(B)=41.2470$ & $\sigma^2_S(B)=37.1640$

Equating equation: $l(b)=.9367(b-41.2470)+41.1826$

Note:

A= Test Form "Book-A" & $_A$= Population taking Book-A
B= Test Form "Book-B" & $_B$= Population taking Book-B
V= Common Items
$_S$= the Synthetic Population
W_A= the Weight for Population A = .4873
W_B= the Weight for Population B = .5127
α= regression coefficient
b= the observed score on Book-B

Accuracy of Equating

The examinees' total raw scores on the 145 common items from the complete test forms were computed and used as the "true score" to study the accuracy of equating, because all the anchors in the reduced test forms were sampled from the 145 common items. The IRT true score estimates yielded by the two IRT equating methods, as well as the scaled scores obtained by Tucker's linear equating, for all the reduced test forms were correlated to the "true score". The correlation analyses were summarized in TABLE 4.

IRT True Score Estimates & Tucker's Scaled Scores

Overall, the results of the IRT two-stage and fixed-b equating were moderately accurate. The correlation coefficients between the "true scores" and the true score estimates ranged from .832 to .856, over the three sets of reduced forms. The correlation coefficients for Tucker's linear equating ranged from .802 to .832, which also indicated moderate accuracy. The correlation analyses showed that the equating, regardless of the equating method, generally ordered the examinees in a way consistent to their standings based on their true scores.

Although the magnitude of the correlation coefficients were similar, the IRT equating always had higher correlation coefficients. It seemed that the two IRT equating methods always resulted in more accurate scores than Tucker's linear equating. TABLE 4 also showed that the true score estimates yielded by the two IRT methods correlated almost perfectly, .999, over various

TABLE 4: Correlation Analysis

–Pearson Coefficients of Correlation Between the 'True Scores' and Various True Score Estimates

		"True Score"	IRT Two-stage Equating			IRT Fixed-b Equating			Tucker's Linear Equating		
			12 Anchors	20 Anchors	30 Anchors	12 Anchors	20 Anchors	30 Anchors	12 Anchors	20 Anchors	30 Anchors
"True Score"		1.0000 0.0000	0.8318 0.0001	0.8472 0.0001	0.8556 0.0001	0.8318 0.0001	0.8469 0.0001	0.8541 0.0001	0.8017 0.0001	0.8230 0.0001	0.8319 0.0001
IRT Two-stage Equating	12 Anchors	0.8318 0.0001	1.0000 0.0000	0.9774 0.0001	0.9270 0.0001	0.9993 0.0001	0.9769 0.0001	0.9258 0.0001	0.9590 0.0001	0.9339 0.0001	0.8856 0.0001
	20 Anchors	0.8472 0.0001	0.9774 0.0001	1.0000 0.0000	0.9470 0.0001	0.9779 0.0001	0.9996 0.0001	0.9450 0.0001	0.9307 0.0001	0.9616 0.0001	0.9069 0.0001
	30 Anchors	0.8556 0.0001	0.9270 0.0001	0.9470 0.0001	1.0000 0.0000	0.9266 0.0001	0.9468 0.0001	0.9994 0.0001	0.8918 0.0001	0.9186 0.0001	0.9645 0.0001
IRT Fixed-b Equating	12 Anchors	0.8318 0.0001	0.9993 0.0001	0.9779 0.0001	0.9266 0.0001	1.0000 0.0000	0.9778 0.0001	0.9249 0.0001	0.9571 0.0001	0.9329 0.0001	0.8845 0.0001
	20 Anchors	0.8469 0.0001	0.9769 0.0001	0.9996 0.0001	0.9468 0.0001	0.9778 0.0001	1.0000 0.0000	0.9454 0.0001	0.9297 0.0001	0.9610 0.0001	0.9065 0.0001
	30 Anchors	0.8541 0.0001	0.9258 0.0001	0.9450 0.0001	0.9994 0.0001	0.9249 0.0001	0.9454 0.0001	1.0000 0.0000	0.8903 0.0001	0.9177 0.0001	0.9634 0.0001
Tucker's Linear Equating	12 Anchors	0.8017 0.0001	0.9590 0.0001	0.9307 0.0001	0.8918 0.0001	0.9571 0.0001	0.9297 0.0001	0.8903 0.0001	1.0000 0.0000	0.9525 0.0001	0.9012 0.0001
	20 Anchors	0.8230 0.0001	0.9339 0.0001	0.9616 0.0001	0.9186 0.0001	0.9329 0.0001	0.9610 0.0001	0.9177 0.0001	0.9525 0.0001	1.0000 0.0000	0.9314 0.0001
	30 Anchors	0.8319 0.0001	0.8856 0.0001	0.9069 0.0001	0.9645 0.0001	0.8845 0.0001	0.9065 0.0001	0.9634 0.0001	0.9012 0.0001	0.9314 0.0001	1.0000 0.0000

Note: Prob > |R| under Ho: Rho=0
N=2241

reduced forms. It was then concluded that there was no difference in the results of the two IRT equating methods.

Anchor Effect

For IRT two-stage equating, the correlation coefficient increased from .832 to .847 to .856, when the number of anchor items increased from 12 to 20 to 30. For IRT fixed-b equating, the coefficient increased from .832 to .847 to .854, and for Tucker's equating it increased from .802 to .823 to .832. The patterns generally suggested that equating would be improved given more anchor items, no matter which of the three equating methods was used. Nevertheless, it should be noted that the improvement might not be practically significant.

The above findings suggested that, if both the sufficiency and efficiency of an anchor were concerned in equating tests, test forms that had at least 20 anchor items or 12 items (1/5 of the total test in length) were desirable. In summary, anchor length seemed to have substantial impacts on test equating, regardless of the method of equating. It was important to include enough number of anchor items in equating practice.

IRT Estimates For Unique Items

There was a concern that the correlation between the true scores and the true score estimates based on the entire reduced forms might have been inflated by auto-correlation. Therefore, the "true scores" were also correlated with the IRT true score estimates for various reduced forms that did not include any anchor items (see APPENDIX 5). The same correlation analysis,

however, was not done for Tucker's linear equating results because the method was based on the observed test scores as a whole. In addition, the scaled scores based on only the unique items could not be computed for Tucker's linear equating results. It had been shown that IRT methods were more flexible in computing the scaled scores, after equating, if certain items were added or dropped.

The auto-correlation of concern was due to the overlapped anchor items in the reduced forms and the complete 145 anchor set. By correlating the true scores with the true score estimates using unique items only, the auto-correlation still could not be totally eliminated. It was because the IRT estimation for the unique items was still influenced by the characteristics of the anchors, especially for the study using common-item design. However, by controlling for part of the auto-correlation, the correlation analysis results presented in APPENDIX 5 at least provided a better picture to for understanding the goodness of different equating methods.

The patterns of the correlation coefficients, discovered in the previous section of anchor effect, were expected to have some changes. It was because different numbers of anchor items (12, 20, or 30) were excluded from the correlation analysis. The results in APPENDIX 5 showed that, despite the slight decrease in the magnitude of the coefficients, the general patterns did not change. Therefore, the anchor length effect and the similarity of the two IRT equating methods, in terms of the equating results,

were retained.

Another possible explanation for the unchanged patterns was that the anchor and the unique items in the three pairs of reduced forms were homogeneous. As a result, the differences in the correlation coefficients were minimal when the anchor items were excluded. The correlation analysis for the raw scores had a similar results. Furthermore, the decrease in the magnitude of the coefficients could be attributed to the smaller number of items used in computing the correlation coefficients, after excluding 12, 20, or 30 common items.

IRT Estimates For Anchor Items

To further investigate the goodness of IRT equating, as well as the reliability of the anchor, the "true scores" were also correlated with the IRT true score estimates obtained by using the anchor items only. The results of the correlation analysis was summarized in APPENDIX 6.

As before, the patterns of the coefficients still suggested substantial effect of anchor length. The longer the anchor length was, the better the equating result would be. The magnitude of the correlation coefficients did not change much, after excluding the unique items from the correlation analysis. In addition, the results of the two IRT equating methods were close as discovered previously. As before, the reduction in the magnitude of the coefficients might due to the smaller number of items used in computing the coefficients.

In relation to the true score estimates based on the anchor items only, the "true score" obtained from the 145 anchor items was regarded as a similar but more reliable measure. It was because the "true score" was computed using more items. As a consequence, the "true score" could be used as the criterion measure to study the concurrent validity of the anchor. Generally, the high correlation coefficients between the "true score" and various true score estimates, shown in APPENDIX 6, provided evidence of high validity for the three anchors. In average, the coefficient for the 12-item anchor was .832. For the 20-item anchor, it was .847. And for the 30-item anchor, it was higher, .856.

From the perspective of correlating an ability measure and its corresponding true score, the coefficients in APPENDIX 6 might also be regarded as reliability measures. The coefficients, then, suggested that the anchors ware considerably reliable. Given the adequate validity and reliability, along with the more accurate equating results, both IRT two-stage method and fixed-b method were considered satisfactory.

Critiques on the Criteria of Equating Accuracy

The "true score" obtained from the complete 145 anchor items was used as a criterion for equating accuracy, because it could be regarded as the item population from which the common items in the pairs of the reduced forms were drawn. Nevertheless, this

criterion was only appropriate when the examinee population and the testing occasion were fixed.

Despite all the nice features of being longer and thus more reliable, taken by both of the examinee groups, as well as the content similarity to the forms being equated, the criterion was at most a convenient but close approximation to the true score. Because it was the raw-score total of the zero/one (wrong/right) coded items, it could not escape from the common drawbacks of raw-score-based measures. For instance, the characteristics of the items were not "person-free" and the scores were not "item-free".

Nevertheless, the 145 items constituted a conceptually reasonable item population, and all the items were taken by the entire examinee population. As a result, this raw-score-based criterion was blameless for not being "person-free" or "item-free".

Alternatively, an IRT estimated score could be computed using the 145 common items to serve as a criterion of equating accuracy. Nonetheless, the IRT-based criterion might be biased in favor of the IRT equating methods. For Tucker's linear equating, which was quite different for IRT equating, the IRT-based criterion could underestimate its equating accuracy. Taking into account the issues, the raw-score-based criterion was used in this study to obtain a conservative estimate of equating accuracy for IRT equating.

BEST COPY AVAILABLE

Adequacy of the Guessing Parameter

With the above findings on item calibration and equating accuracy, it was concluded that the three parameter IRT model fit the minimum competence test data that was used in this study. The inclusion of the guessing parameter was reasonable because of the chance of guessing on difficult items, due to the nature of the multiple-choice format, as well as the strong motive of the examinees to obtain higher scores. It was also justified by the empirical results of equating. Therefore, it seemed appropriate to include the guessing parameter when equating tests or test forms that had negatively skewed score distributions.

Suggestions

In addition to the current research, another study can be done to investigate the function of various equating methods, when the test forms become longer or the number of anchor items are increased. If the test data of different years are available, cross-year equating can be done to study the effects of test and examinee characteristics over time. Validation study could be conducted to further determine the adequacy of equating, if the examinees' performances on the formal licensure exams were available.

Equating accuracy could be better studied if certain unbiased criteria were identified. Due to the restrictions on the current research design and the nature of the test items, test forms that had very short but adequate anchors were not studied. If such

short anchors were available in the future, efficiency of the anchor could be further studied. Only the three parameter logistic IRT model was applied in this study, because of the possibility of examinee's guessing on the difficult items. However, for a minimum competency test, if guessing is not a major concern, the two parameter IRT model or the Rasch model might as well fit the data. Further investigation are needed for the fit of different IRT models to reduce the cost in the actual equating process.

For test forms used in personnel selection or certification, usually a cut-off score is arbitrarily established. It will be useful to know how the results of equating function with the arbitrary cut-off standard. For example, it would be interesting to learn about the influence of the IRT equating on the hit and miss rates, under the impact of an arbitrary criterion such as "the top 75% of the examinees".

An interesting variation of the current study is to examine the effect of content mix and equating method on the accuracy of test equating by comparing the linear and IRT Equating results, following the same anchor item design of this study. The manipulation of the content of the test forms is reasonable because the 255 test items falls into 23 content sub-areas. Assuming substantial differences among the content areas, from the big item pool, items could be sampled by different schemes to obtain reduced test forms that are different in terms of content mix. In a separate study, four pairs of shorter forms were

created using various sampling schemes, including simple random sampling, equal weight domain random sampling, proportional weight domain random sampling, and purposeful sampling. The results supported that equating accuracy depended on the content representativeness of the anchor items.

Reference

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, N.J.: Educational Testing Service.

Baker, F. B. (1990). Some observations o the metric of PC-BILOG results. Applied psychological measurement, 14, 139-150.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. Journal of educational measurement, 28, 147-162.

Berk, R. H. (1982). Discussion of item response theory. In P. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. Applied psychological measurement, 11, 279-290.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of educational measurement, 22, 13-20.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory (pp.175-195). Vancouver, British Columbia: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. Educational measurement: Issues and Practice, 10, 37-45.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied psychological measurement, 11, 225-244.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1988). The effects on IRT and conventional achievement test equating results of using equating samples matched on ability (Research Rep. No. RR-88 -52). Princeton, NJ: Educational Testing Service.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Chicago: Holt, Rinehart and Winston, Inc.

Dorans, N. J. (1990). Equating methods and sampling designs. Applied measurement in education, 3, 3-17.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. Journal of educational measurement, 22, 249-262.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true -score equating procedures of matching on a fallible criterion. Applied measurement in education, 3, 37-52.

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. Applied measurement in education, 2, 297-312.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of educational measurement, 14, 75-96.

Hambleton, R. K., & Swaminathan, H. (1990). Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage.

Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum-competency tests: Comparisons of methods. Journal of educational measurement, 25, 221-231.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. Journal of educational measurement, 22, 197-206.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. Journal of educational measurement, 18, 1-10.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. Applied psychological measurement, 11, 263-277.

Kolen, M. J., & Harris, D. J. (1990). Comparison of Item preequating and random groups equating using IRT and equipercentile methods. Journal of educational measurement, 27, 27-39.

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. Applied measurement in education, 3, 19-36.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? Applied measurement in education, 3, 73-95.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of educational measurement, 14, 117-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1982). Item response theory and equating -- A technical summary. In P. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Marco, G. L., Petersen, N. C., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. J. Weiss (Ed.), New horizons in testing: Latent trait theory and computerized adaptive testing. New York: Academic Press.

Mislevy, R. J., & Bock, R. D. (1990). Bilog 3: Item analysis and test scoring with binary logistic models. Mooresville, I.N.: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied psychological measurement, 13, 57-75.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of educational statistics, 8, 137-156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, Norming, and Equating. In R. L. Linn (Ed.), Educational Measurement. New York: ACE/Macmillan.

Raju, N. S., Bode, R. K., Larsen, V. S., & Steinhaus, S. (1986, April). Anchor-test size and horizontal equating with the Rasch and three-parameter models. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983, April). The effect of anchor test size in vertical equating with the Rasch and three-parameter models. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational measurement, 25, 193-203.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. Applied Measurement in Education, 3, 53-71.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of educational research, 56, 495-529.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. Applied psychological measurement, 12, 69-82.

Stocking, M. L., & Lord, F. M. (1982). Developing a common metric in item response theory (Research report RR-82-5-ONR). Princeton, N.J.: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied psychological measurement, 8, 347-364.

Wingersky, M. S., & Barton, M. A. (1982). Logist user's guide: Logist 5, Version 1.0. Princeton, N.J.: Educational Testing Service.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of educational measurement, 17, 297-311.

BEST COPY AVAILABLE

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied psychological measurement, 8, 125-145.

Yen, W. M. (1985). Tau equivalence of vertical equating using three-parameter item response theory and Thurstonian procedures. Paper presented at the meeting of the American Educational Research Association, Chicago.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

Appendix 1: Item Sampling Scheme for the Reduced Forms

Simple Random Sampling

Assumption: There is no substantial differences among the items from the 23 content sub-areas, since all the items are from a single item pool for emergency medicine.

Method: Mix the items from the 23 content areas and randomly sample from the pool using a random number table.

Results: Three pairs of reduced test forms.
There are 60 items in each form.

Special design:

The control for the anchor length is incorporated with the simple random sampling of the test items to study the effect of the anchor length.

Taking into account the recommended anchor lengths from previous research findings, three samples consisting 30, 20, and 12 anchor items are drawn respectively.

The scenarios of the recommendation for the appropriate anchor length are as follows:

1. at least 20 anchor items, or
2. at least 20% of the total test (Angoff, 1984); and
3. 5 or 6 carefully chosen items may suffice.

Results: Three Pairs of Reduced Test Forms

Items randomly selected for the samples of--

30 anchor and 30 unique items:
Book-A: Length=60 (items)
Book-B: Length=60

20 anchor and 40 unique items:
Book-A: Length=60
Book-B: Length=60

12 anchor and 48 unique items:
Book-A: LENGTH=60
Book-B: LENGTH=60

Appendix 2: Item Difficulty & Item-Total Correlation for the Items in the Reduced Test Forms

Simple Random Sample with 30 Anchor Items

CLASSICAL ITEM STATISTICS FOR SUBTEST Book-A

ITEM	NAME	NUMBER TRIED	NUMBER RIGHT	PERCENT	LOGIT/1.7	ITEM*TEST CORRELATION PEARSON	BISERIAL
1	0001	1092.0	776.0	.711	.53	.109	.144
2	0002	1092.0	1016.0	.930	1.53	.167	.318
3	0003	1092.0	813.0	.745	.63	.170	.231
4	0004	1092.0	998.0	.914	1.39	.220	.392
5	0005	1092.0	826.0	.756	.67	-.020	-.027
6	0006	1092.0	945.0	.865	1.09	.091	.143
7	0007	1092.0	554.0	.507	.02	.080	.101
8	0008	1092.0	527.0	.483	-.04	.311	.390
9	0009	1092.0	839.0	.768	.71	-.001	-.001
10	0010	1092.0	982.0	.899	1.29	.144	.246
11	0011	1092.0	699.0	.640	.34	.003	.004
12	0012	1092.0	950.0	.870	1.12	.074	.117
13	0013	1092.0	456.0	.418	-.20	.109	.137
14	0014	1092.0	883.0	.809	.85	.353	.509
15	0015	1092.0	907.0	.831	.94	.099	.147
16	0016	1092.0	928.0	.850	1.02	.192	.294
17	0017	1092.0	718.0	.658	.38	.281	.363
18	0018	1092.0	768.0	.703	.51	.075	.099
19	0019	1092.0	840.0	.769	.71	.198	.274
20	0020	1092.0	727.0	.666	.41	.108	.139
21	0021	1092.0	655.0	.600	.24	.135	.171
22	0022	1092.0	795.0	.728	.58	.133	.178
23	0023	1092.0	785.0	.719	.55	.112	.150
24	0024	1092.0	832.0	.762	.68	.131	.180
25	0025	1092.0	605.0	.554	.13	.120	.151
26	0026	1092.0	970.0	.888	1.22	.207	.344
27	0027	1092.0	534.0	.489	-.03	.191	.239
28	0028	1092.0	973.0	.891	1.24	.110	.184
29	0029	1092.0	531.0	.486	-.03	.285	.357
30	0030	1092.0	826.0	.756	.67	.204	.279
31	0031	1092.0	720.0	.659	.39	.208	.269
32	0032	1092.0	961.0	.880	1.17	.043	.070
33	0033	1092.0	501.0	.459	-.10	.137	.171
34	0034	1092.0	544.0	.498	.00	.181	.226
35	0035	1092.0	743.0	.680	.44	.127	.166
36	0036	1092.0	769.0	.704	.51	.098	.130
37	0037	1092.0	802.0	.734	.60	.085	.114
38	0038	1092.0	1000.0	.916	1.40	.153	.276
39	0039	1092.0	738.0	.676	.43	.005	.006
40	0040	1092.0	610.0	.559	.14	-.044	-.055
41	0041	1092.0	716.0	.656	.38	.309	.399
42	0042	1092.0	1015.0	.929	1.52	.088	.167
43	0043	1092.0	656.0	.601	.24	.306	.388
44	0044	1092.0	658.0	.603	.24	.201	.255
45	0045	1092.0	801.0	.734	.60	.100	.135
46	0046	1092.0	748.0	.685	.46	.230	.301
47	0047	1092.0	369.0	.338	-.40	.147	.190
48	0048	1092.0	613.0	.561	.15	.041	.052
49	0049	1092.0	1060.0	.971	2.06	.082	.205
50	0050	1092.0	720.0	.659	.39	.102	.132
51	0051	1092.0	713.0	.653	.37	.170	.220
52	0052	1092.0	971.0	.889	1.23	.263	.437
53	0053	1092.0	470.0	.430	-.16	.177	.223
54	0054	1092.0	584.0	.535	.08	.096	.121
55	0055	1092.0	528.0	.484	-.04	.128	.160
56	0056	1092.0	635.0	.582	.19	.152	.192
57	0057	1092.0	559.0	.512	.03	.121	.151
58	0058	1092.0	764.0	.700	.50	.137	.181
59	0059	1092.0	795.0	.728	.58	.163	.218
60	0060	1092.0	661.0	.605	.25	.141	.179

Simple Random Sample with 30 Anchor Items

CLASSICAL ITEM STATISTICS FOR SUBTEST Book-B

ITEM	NAME	NUMBER TRIED	NUMBER RIGHT	PERCENT	LOGIT/1.7	ITEM*TEST CORRELATION PEARSON	BISERIAL
1	0001	1149.0	861.0	.749	.64	.101	.137
2	0002	1149.0	1085.0	.944	1.66	.148	.302
3	0003	1149.0	898.0	.782	.75	.142	.199
4	0004	1149.0	1038.0	.903	1.32	.252	.435
5	0005	1149.0	901.0	.784	.76	.031	.044
6	0006	1149.0	998.0	.869	1.11	.150	.237
7	0007	1149.0	595.0	.518	.04	.132	.165
8	0008	1149.0	591.0	.514	.03	.217	.272
9	0009	1149.0	872.0	.759	.67	.006	.008
10	0010	1149.0	1047.0	.911	1.37	.148	.262
11	0011	1149.0	722.0	.628	.31	.123	.157
12	0012	1149.0	984.0	.856	1.05	.060	.093
13	0013	1149.0	495.0	.431	-.16	.202	.255
14	0014	1149.0	998.0	.869	1.11	.333	.528
15	0015	1149.0	987.0	.859	1.06	.123	.192
16	0016	1149.0	1006.0	.876	1.15	.158	.254
17	0017	1149.0	786.0	.684	.45	.351	.459
18	0018	1149.0	762.0	.663	.40	.053	.069
19	0019	1149.0	913.0	.795	.80	.173	.245
20	0020	1149.0	789.0	.687	.46	.132	.173
21	0021	1149.0	714.0	.621	.29	.167	.213
22	0022	1149.0	810.0	.705	.51	.194	.256
23	0023	1149.0	859.0	.748	.64	.210	.286
24	0024	1149.0	876.0	.762	.69	.136	.188
25	0025	1149.0	622.0	.541	.10	.063	.079
26	0026	1149.0	999.0	.869	1.12	.273	.434
27	0027	1149.0	612.0	.533	.08	.192	.241
28	0028	1149.0	1034.0	.900	1.29	.163	.279
29	0029	1149.0	609.0	.530	.07	.285	.358
30	0030	1149.0	917.0	.798	.81	.165	.235
31	0031	1149.0	907.0	.789	.78	.229	.324
32	0032	1149.0	736.0	.641	.34	.182	.234
33	0033	1149.0	633.0	.551	.12	.211	.265
34	0034	1149.0	621.0	.540	.10	.274	.344
35	0035	1149.0	623.0	.542	.10	.160	.201
36	0036	1149.0	861.0	.749	.64	.149	.202
37	0037	1149.0	801.0	.697	.49	.175	.230
38	0038	1149.0	946.0	.823	.91	.192	.282
39	0039	1149.0	594.0	.517	.04	.207	.260
40	0040	1149.0	562.0	.489	-.03	.256	.321
41	0041	1149.0	544.0	.473	-.06	.157	.197
42	0042	1149.0	958.0	.834	.95	.054	.081
43	0043	1149.0	951.0	.828	.92	.289	.427
44	0044	1149.0	724.0	.630	.31	.252	.322
45	0045	1149.0	1017.0	.885	1.20	.124	.204
46	0046	1149.0	677.0	.589	.21	.043	.054
47	0047	1149.0	1032.0	.898	1.28	.126	.215
48	0048	1149.0	914.0	.795	.80	.192	.273
49	0049	1149.0	538.0	.468	-.07	.063	.079
50	0050	1149.0	469.0	.408	-.22	.160	.203
51	0051	1149.0	993.0	.864	1.09	.277	.435
52	0052	1149.0	711.0	.619	.28	.336	.429
53	0053	1149.0	511.0	.445	-.13	.142	.179
54	0054	1149.0	939.0	.817	.88	.090	.131
55	0055	1149.0	536.0	.466	-.08	.111	.139
56	0056	1149.0	1054.0	.917	1.42	.190	.344
57	0057	1149.0	672.0	.585	.20	.051	.064
58	0058	1149.0	912.0	.794	.79	.085	.121
59	0059	1149.0	803.0	.699	.50	.237	.312
60	0060	1149.0	850.0	.740	.61	.164	.221

Simple Random Sample with 20 Anchor Items

CLASSICAL ITEM STATISTICS FOR SUBTEST Book-A

ITEM	NAME	NUMBER TRIED	NUMBER RIGHT	PERCENT	LOGIT/1.7	ITEM*TEST CORRELATION PEARSON	BISERIAL
1	0001	1092.0	776.0	.711	.53	.104	.138
2	0002	1092.0	813.0	.745	.63	.163	.222
3	0003	1092.0	998.0	.914	1.39	.228	.407
4	0004	1092.0	826.0	.756	.67	-.031	-.043
5	0005	1092.0	554.0	.507	.02	.090	.112
6	0006	1092.0	527.0	.483	-.04	.336	.421
7	0007	1092.0	982.0	.899	1.29	.154	.263
8	0008	1092.0	699.0	.640	.34	.009	.011
9	0009	1092.0	456.0	.418	-.20	.098	.124
10	0010	1092.0	883.0	.809	.85	.354	.511
11	0011	1092.0	928.0	.850	1.02	.212	.324
12	0012	1092.0	718.0	.658	.38	.292	.377
13	0013	1092.0	840.0	.769	.71	.219	.303
14	0014	1092.0	727.0	.666	.41	.108	.141
15	0015	1092.0	655.0	.600	.24	.140	.177
16	0016	1092.0	795.0	.728	.58	.122	.163
17	0017	1092.0	832.0	.762	.68	.139	.192
18	0018	1092.0	605.0	.554	.13	.101	.127
19	0019	1092.0	534.0	.489	-.03	.216	.271
20	0020	1092.0	826.0	.756	.67	.211	.289
21	0021	1092.0	632.0	.579	.19	.174	.220
22	0022	1092.0	720.0	.659	.39	.212	.274
23	0023	1092.0	702.0	.643	.35	.115	.148
24	0024	1092.0	961.0	.880	1.17	.042	.068
25	0025	1092.0	501.0	.459	-.10	.147	.185
26	0026	1092.0	544.0	.498	.00	.185	.232
27	0027	1092.0	743.0	.680	.44	.133	.173
28	0028	1092.0	769.0	.704	.51	.110	.146
29	0029	1092.0	802.0	.734	.60	.070	.094
30	0030	1092.0	1000.0	.916	1.40	.168	.303
31	0031	1092.0	738.0	.676	.43	-.005	-.006
32	0032	1092.0	866.0	.793	.79	.345	.489
33	0033	1092.0	610.0	.559	.14	-.021	-.027
34	0034	1092.0	716.0	.656	.38	.324	.418
35	0035	1092.0	1015.0	.929	1.52	.082	.156
36	0036	1092.0	656.0	.601	.24	.309	.391
37	0037	1092.0	658.0	.603	.24	.187	.237
38	0038	1092.0	800.0	.733	.59	.229	.308
39	0039	1092.0	801.0	.734	.60	.089	.119
40	0040	1092.0	770.0	.705	.51	.149	.197
41	0041	1092.0	748.0	.685	.46	.228	.299
42	0042	1092.0	369.0	.338	-.40	.149	.193
43	0043	1092.0	613.0	.561	.15	.036	.045
44	0044	1092.0	738.0	.676	.43	.028	.036
45	0045	1092.0	925.0	.847	1.01	.053	.081
46	0046	1092.0	952.0	.872	1.13	.218	.349
47	0047	1092.0	1060.0	.971	2.06	.074	.184
48	0048	1092.0	720.0	.659	.39	.099	.128
49	0049	1092.0	829.0	.759	.68	.253	.347
50	0050	1092.0	971.0	.889	1.23	.270	.449
51	0051	1092.0	470.0	.430	-.16	.179	.226
52	0052	1092.0	427.0	.391	-.26	-.010	-.013
53	0053	1092.0	627.0	.574	.18	.186	.234
54	0054	1092.0	528.0	.484	-.04	.141	.176
55	0055	1092.0	635.0	.582	.19	.138	.175
56	0056	1092.0	1004.0	.919	1.43	.172	.313
57	0057	1092.0	559.0	.512	.03	.132	.165
58	0058	1092.0	764.0	.700	.50	.157	.207
59	0059	1092.0	913.0	.836	.96	.187	.280
60	0060	1092.0	661.0	.605	.25	.151	.192

Simple Random Sample with 20 Anchor Items

CLASSICAL ITEM STATISTICS FOR SUBTEST Book-B

ITEM	NAME	NUMBER TRIED	NUMBER RIGHT	PERCENT	LOGIT/1.7	ITEM*TEST CORRELATION PEARSON	BISERIAL
1	0001	1149.0	861.0	.749	.64	.096	.131
2	0002	1149.0	898.0	.782	.75	.151	.212
3	0003	1149.0	1038.0	.903	1.32	.235	.405
4	0004	1149.0	901.0	.784	.76	.019	.027
5	0005	1149.0	595.0	.518	.04	.123	.154
6	0006	1149.0	591.0	.514	.03	.232	.290
7	0007	1149.0	1047.0	.911	1.37	.142	.251
8	0008	1149.0	722.0	.628	.31	.116	.149
9	0009	1149.0	495.0	.431	-.16	.186	.235
10	0010	1149.0	998.0	.869	1.11	.333	.527
11	0011	1149.0	1006.0	.876	1.15	.173	.278
12	0012	1149.0	786.0	.684	.45	.361	.471
13	0013	1149.0	913.0	.795	.80	.164	.233
14	0014	1149.0	789.0	.687	.46	.138	.181
15	0015	1149.0	714.0	.621	.29	.148	.189
16	0016	1149.0	810.0	.705	.51	.206	.272
17	0017	1149.0	876.0	.762	.69	.133	.183
18	0018	1149.0	622.0	.541	.10	.078	.098
19	0019	1149.0	612.0	.533	.08	.176	.221
20	0020	1149.0	917.0	.798	.81	.181	.257
21	0021	1149.0	907.0	.789	.78	.250	.354
22	0022	1149.0	736.0	.641	.34	.177	.228
23	0023	1149.0	633.0	.551	.12	.209	.263
24	0024	1149.0	621.0	.540	.10	.280	.351
25	0025	1149.0	623.0	.542	.10	.164	.206
26	0026	1149.0	643.0	.560	.14	-.007	-.009
27	0027	1149.0	861.0	.749	.64	.160	.219
28	0028	1149.0	801.0	.697	.49	.182	.240
29	0029	1149.0	894.0	.778	.74	.036	.050
30	0030	1149.0	646.0	.562	.15	.179	.225
31	0031	1149.0	946.0	.823	.91	.212	.312
32	0032	1149.0	594.0	.517	.04	.202	.253
33	0033	1149.0	562.0	.489	-.03	.247	.310
34	0034	1149.0	966.0	.841	.98	.189	.286
35	0035	1149.0	544.0	.473	-.06	.146	.183
36	0036	1149.0	986.0	.858	1.06	.257	.399
37	0037	1149.0	951.0	.828	.92	.298	.442
38	0038	1149.0	724.0	.630	.31	.258	.330
39	0039	1149.0	1017.0	.885	1.20	.139	.229
40	0040	1149.0	679.0	.591	.22	.045	.057
41	0041	1149.0	565.0	.492	-.02	.005	.006
42	0042	1149.0	995.0	.866	1.10	.280	.442
43	0043	1149.0	563.0	.490	-.02	.218	.273
44	0044	1149.0	726.0	.632	.32	.228	.292
45	0045	1149.0	1032.0	.898	1.28	.133	.226
46	0046	1149.0	914.0	.795	.80	.172	.245
47	0047	1149.0	469.0	.408	-.22	.164	.207
48	0048	1149.0	1111.0	.967	1.99	.232	.559
49	0049	1149.0	993.0	.864	1.09	.271	.426
50	0050	1149.0	711.0	.619	.28	.342	.436
51	0051	1149.0	939.0	.817	.88	.091	.133
52	0052	1149.0	922.0	.802	.82	-.001	-.001
53	0053	1149.0	536.0	.466	-.08	.108	.136
54	0054	1149.0	942.0	.820	.89	.217	.318
55	0055	1149.0	1054.0	.917	1.42	.175	.316
56	0056	1149.0	672.0	.585	.20	.056	.070
57	0057	1149.0	934.0	.813	.86	.121	.176
58	0058	1149.0	912.0	.794	.79	.059	.084
59	0059	1149.0	1025.0	.892	1.24	.322	.539
60	0060	1149.0	850.0	.740	.61	.179	.242

Simple Random Sample with 12 Anchor Items

CLASSICAL ITEM STATISTICS FOR SUBTEST Book-A

ITEM	NAME	NUMBER TRIED	NUMBER RIGHT	PERCENT	LOGIT/1.7	ITEM*TEST CORRELATION PEARSON	BISERIAL
1	0001	1092.0	776.0	.711	.53	.115	.152
2	0002	1092.0	813.0	.745	.63	.168	.228
3	0003	1092.0	554.0	.507	.02	.099	.124
4	0004	1092.0	527.0	.483	-.04	.330	.414
5	0005	1092.0	982.0	.899	1.29	.154	.263
6	0006	1092.0	883.0	.809	.85	.353	.509
7	0007	1092.0	718.0	.658	.38	.289	.373
8	0008	1092.0	727.0	.666	.41	.114	.148
9	0009	1092.0	795.0	.728	.58	.121	.162
10	0010	1092.0	605.0	.554	.13	.089	.112
11	0011	1092.0	534.0	.489	-.03	.215	.269
12	0012	1092.0	826.0	.756	.67	.220	.301
13	0013	1092.0	632.0	.579	.19	.166	.210
14	0014	1092.0	720.0	.659	.39	.202	.262
15	0015	1092.0	702.0	.643	.35	.103	.133
16	0016	1092.0	961.0	.880	1.17	.034	.056
17	0017	1092.0	501.0	.459	-.10	.141	.177
18	0018	1092.0	544.0	.498	.00	.164	.206
19	0019	1092.0	743.0	.680	.44	.137	.179
20	0020	1092.0	769.0	.704	.51	.112	.148
21	0021	1092.0	802.0	.734	.60	.071	.095
22	0022	1092.0	1000.0	.916	1.40	.158	.283
23	0023	1092.0	900.0	.824	.91	.037	.054
24	0024	1092.0	738.0	.676	.43	-.005	-.006
25	0025	1092.0	866.0	.793	.79	.357	.506
26	0026	1092.0	610.0	.559	.14	-.034	-.043
27	0027	1092.0	716.0	.656	.38	.317	.409
28	0028	1092.0	1015.0	.929	1.52	.080	.152
29	0029	1092.0	656.0	.601	.24	.296	.375
30	0030	1092.0	1036.0	.949	1.72	.078	.162
31	0031	1092.0	658.0	.603	.24	.187	.237
32	0032	1092.0	800.0	.733	.59	.204	.274
33	0033	1092.0	801.0	.734	.60	.087	.117
34	0034	1092.0	770.0	.705	.51	.160	.211
35	0035	1092.0	748.0	.685	.46	.223	.292
36	0036	1092.0	369.0	.338	-.40	.152	.197
37	0037	1092.0	613.0	.561	.15	.052	.066
38	0038	1092.0	812.0	.744	.63	.070	.095
39	0039	1092.0	738.0	.676	.43	.016	.020
40	0040	1092.0	925.0	.847	1.01	.068	.104
41	0041	1092.0	952.0	.872	1.13	.213	.340
42	0042	1092.0	1060.0	.971	2.06	.082	.205
43	0043	1092.0	720.0	.659	.39	.096	.124
44	0044	1092.0	1003.0	.918	1.42	.111	.202
45	0045	1092.0	829.0	.759	.68	.252	.346
46	0046	1092.0	773.0	.708	.52	-.011	-.014
47	0047	1092.0	971.0	.889	1.23	.271	.449
48	0048	1092.0	470.0	.430	-.16	.198	.250
49	0049	1092.0	427.0	.391	-.26	-.011	-.013
50	0050	1092.0	627.0	.574	.18	.173	.218
51	0051	1092.0	528.0	.484	-.04	.132	.165
52	0052	1092.0	635.0	.582	.19	.134	.169
53	0053	1092.0	1004.0	.919	1.43	.178	.323
54	0054	1092.0	559.0	.512	.03	.149	.187
55	0055	1092.0	465.0	.426	-.18	.119	.150
56	0056	1092.0	975.0	.893	1.25	.104	.174
57	0057	1092.0	764.0	.700	.50	.154	.203
58	0058	1092.0	913.0	.836	.96	.200	.300
59	0059	1092.0	404.0	.370	-.31	.236	.302
60	0060	1092.0	661.0	.605	.25	.163	.207

Simple Random Sample with 12 Anchor Items

CLASSICAL ITEM STATISTICS FOR SUBTEST Book-B

ITEM	NAME	NUMBER TRIED	NUMBER RIGHT	PERCENT	LOGIT/1.7	ITEM*TEST CORRELATION PEARSON	BISERIAL
1	0001	1149.0	861.0	.749	.64	.083	.112
2	0002	1149.0	898.0	.782	.75	.147	.206
3	0003	1149.0	595.0	.518	.04	.121	.151
4	0004	1149.0	591.0	.514	.03	.231	.290
5	0005	1149.0	1047.0	.911	1.37	.154	.273
6	0006	1149.0	998.0	.869	1.11	.318	.504
7	0007	1149.0	786.0	.684	.45	.342	.447
8	0008	1149.0	789.0	.687	.46	.138	.180
9	0009	1149.0	810.0	.705	.51	.190	.251
10	0010	1149.0	622.0	.541	.10	.072	.090
11	0011	1149.0	612.0	.533	.08	.173	.217
12	0012	1149.0	917.0	.798	.81	.180	.257
13	0013	1149.0	907.0	.789	.78	.269	.380
14	0014	1149.0	736.0	.641	.34	.164	.210
15	0015	1149.0	633.0	.551	.12	.205	.258
16	0016	1149.0	621.0	.540	.10	.296	.371
17	0017	1149.0	623.0	.542	.10	.158	.199
18	0018	1149.0	643.0	.560	.14	-.012	-.015
19	0019	1149.0	801.0	.697	.49	.190	.250
20	0020	1149.0	894.0	.778	.74	.043	.060
21	0021	1149.0	646.0	.562	.15	.159	.201
22	0022	1149.0	946.0	.823	.91	.199	.293
23	0023	1149.0	594.0	.517	.04	.201	.252
24	0024	1149.0	562.0	.489	-.03	.237	.297
25	0025	1149.0	966.0	.841	.98	.198	.298
26	0026	1149.0	544.0	.473	-.06	.155	.194
27	0027	1149.0	986.0	.858	1.06	.262	.407
28	0028	1149.0	958.0	.834	.95	.064	.095
29	0029	1149.0	951.0	.828	.92	.301	.445
30	0030	1149.0	724.0	.630	.31	.263	.336
31	0031	1149.0	1017.0	.885	1.20	.134	.221
32	0032	1149.0	679.0	.591	.22	.063	.080
33	0033	1149.0	995.0	.866	1.10	.278	.439
34	0034	1149.0	677.0	.589	.21	.030	.038
35	0035	1149.0	563.0	.490	-.02	.196	.246
36	0036	1149.0	726.0	.632	.32	.245	.313
37	0037	1149.0	1032.0	.898	1.28	.136	.231
38	0038	1149.0	914.0	.795	.80	.179	.255
39	0039	1149.0	610.0	.531	.07	.209	.262
40	0040	1149.0	538.0	.468	-.07	.072	.090
41	0041	1149.0	469.0	.408	-.22	.160	.203
42	0042	1149.0	1111.0	.967	1.99	.235	.565
43	0043	1149.0	993.0	.864	1.09	.278	.436
44	0044	1149.0	711.0	.619	.28	.330	.421
45	0045	1149.0	511.0	.445	-.13	.135	.170
46	0046	1149.0	939.0	.817	.88	.080	.117
47	0047	1149.0	1104.0	.961	1.88	.132	.300
48	0048	1149.0	546.0	.475	-.06	.054	.067
49	0049	1149.0	922.0	.802	.82	-.004	-.006
50	0050	1149.0	536.0	.466	-.08	.093	.116
51	0051	1149.0	1054.0	.917	1.42	.187	.338
52	0052	1149.0	1034.0	.900	1.29	.174	.297
53	0053	1149.0	672.0	.585	.20	.045	.057
54	0054	1149.0	793.0	.690	.47	.108	.142
55	0055	1149.0	934.0	.813	.86	.134	.194
56	0056	1149.0	912.0	.794	.79	.075	.107
57	0057	1149.0	681.0	.593	.22	.213	.269
58	0058	1149.0	1025.0	.892	1.24	.316	.529
59	0059	1149.0	944.0	.822	.90	.090	.131
60	0060	1149.0	850.0	.740	.61	.177	.240

Appendix 3: Descriptive Statistics and Correlation Analyses for the Full Test Forms and the Reduced Test Forms

<Note> A= Items for Book-A B= Items for Book-B
UA= Items unique to Book-A UB= Items unique to Book-B
Z= Common items to both books

All Items In Book-A

Variable	Cases	Mean	Std Dev
A	1092	145.1575	18.0890
Z	1092	105.4570	13.7665
UA	1092	39.7005	5.2857

Variables		Cases	Cross-Prod Dev	Variance-Covar
A	Z	1092	266634.4029	244.3945
A	UA	1092	90352.5055	82.8162
Z	UA	1092	59871.4258	54.8776

- - Correlation Coefficients - -

	A	Z	UA
A	1.0000	.9814**	.8662**
Z	.9814**	1.0000	.7542**
UA	.8662**	.7542**	1.0000

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

All Items in Book-B

Variable	Cases	Mean	Std Dev
B	1149	143.7502	17.3289
Z	1149	107.7206	13.1129
UB	1149	36.0296	5.2337

Variables		Cases	Cross-Prod Dev	Variance-Covar
B	Z	1149	255342.8198	222.4241
B	UB	1149	89392.4926	77.8680
Z	UB	1149	57947.4987	50.4769

- - Correlation Coefficients - -

	B	Z	UB
B	1.0000	.9788**	.8586**
Z	.9788**	1.0000	.7355**
UB	.8586**	.7355**	1.0000

Simple Random Sample with 30 Anchor Items-- For examinees taking Book-A

Variable	Cases	Mean	Std Dev
A	1092	41.2839	5.6591
Z	1092	21.6648	3.1982
UA	1092	19.6190	3.3297

Variables		Cases	Cross-Prod Dev	Variance-Covar
A	Z	1092	17001.9011	15.5838
A	UA	1092	17938.0952	16.4419
Z	UA	1092	5842.5714	5.3552

- - Correlation Coefficients - -

	A	Z	UA
A	1.0000	.8610**	.8726**
Z	.8610**	1.0000	.5029**
UA	.8726**	.5029**	1.0000

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

Simple Random Sample with 30 Anchor Items-- For examinees taking Book-B

Variable	Cases	Mean	Std Dev
B	1149	42.1836	6.0472
Z	1149	22.0888	3.2776
UB	1149	20.0949	3.6187

Variables		Cases	Cross-Prod Dev	Variance-Covar
B	Z	1149	19640.2689	17.1082
B	UB	1149	22339.9835	19.4599
Z	UB	1149	7307.3238	6.3653

- - Correlation Coefficients - -

	B	Z	UB
B	1.0000	.8632**	.8893**
Z	.8632**	1.0000	.5367**
UB	.8893**	.5367**	1.0000

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

Simple Random Sample with 20 Anchor Items-- For examinees taking Book-A

Variable	Cases	Mean	Std Dev
A	1092	40.7427	5.8903
Z	1092	13.7125	2.5710
UA	1092	27.0302	4.1416

Variables		Cases	Cross-Prod Dev	Variance-Covar
A	Z	1092	13175.1996	12.0763
A	UA	1092	24677.4918	22.6191
Z	UA	1092	5963.4890	5.4661

- - Correlation Coefficients - -

	A	Z	UA
A	1.0000	.7974**	.9272**
Z	.7974**	1.0000	.5133**
UA	.9272**	.5133**	1.0000

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

Simple Random Sample with 20 Anchor Items-- For examinees taking Book-B

Variable	Cases	Mean	Std Dev
B	1149	42.1149	6.1097
Z	1149	14.0914	2.5827
UB	1149	28.0235	4.3793

Variables		Cases	Cross-Prod Dev	Variance-Covar
B	Z	1149	14246.9373	12.4102
B	UB	1149	28605.8982	24.9180
Z	UB	1149	6589.5326	5.7400

- - Correlation Coefficients - -

	B	Z	UB
B	1.0000	.7865**	.9313**
Z	.7865**	1.0000	.5075**
UB	.9313**	.5075**	1.0000

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

BEST COPY AVAILABLE

Simple Random Sample with 12 Anchor Items-- For examinees taking Book-A

Variable	Cases	Mean	Std Dev
A	1092	40.8654	5.7666
Z	1092	8.0037	1.9513
UA	1092	32.8617	4.5439

Variables		Cases	Cross-Prod Dev	Variance-Covar
A	Z	1092	8953.5385	8.2067
A	UA	1092	27325.6731	25.0464
Z	UA	1092	4799.5531	4.3992

- - Correlation Coefficients - -

	A	Z	UA
A	1.0000	.7293**	.9559**
Z	.7293**	1.0000	.4962**
UA	.9559**	.4962**	1.0000

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

Simple Random Sample with 12 Anchor Items-- For examinees taking Book-B

Variable	Cases	Mean	Std Dev
B	1149	41.5605	6.0250
Z	1149	8.2907	1.8701
UB	1149	33.2698	4.9109

Variables		Cases	Cross-Prod Dev	Variance-Covar
B	Z	1149	9000.7972	7.8404
B	UB	1149	32672.2489	28.4601
Z	UB	1149	4985.8869	4.3431

- - Correlation Coefficients - -

	B	Z	UB
B	1.0000	.6958**	.9619**
Z	.6958**	1.0000	.4729**
UB	.9619**	.4729**	1.0000

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

APPENDIX 4: Correlation Between Estimates of Ability Scores on Book-B, given by IRT Two-Stage and Fixed-b Equating, for the Three Reduced Item Samples

<Note>
B_ABIL = the ability score yielded by two-stage equating
FB_ABIL= the ability score yielded by fixed-b equating
The estimates on Book-B are **equivalent to the scores on Book-A.**

Simple Random Sample with 30 Common Items

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
B_ABIL	1149	0.1134	0.8687	130.2400	-2.4600	2.5010
FB_ABIL	1149	0.0588	0.8800	67.5220	-2.5650	2.4810

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 1149

	B_ABIL	FB_ABIL
B_ABIL	1.00000	0.99985
	0.0	0.0001
FB_ABIL	0.99985	1.00000
	0.0001	0.0

Simple Random Sample with 20 Common Items

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
B_ABIL	1149	0.1852	0.8764	212.7530	-2.5390	2.6440
FB_ABIL	1149	0.1755	0.9101	201.6550	-2.6070	2.6560

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 1149

	B_ABIL	FB_ABIL
B_ABIL	1.00000	0.99964
	0.0	0.0001
FB_ABIL	0.99964	1.00000
	0.0001	0.0

Simple Random Sample with 12 Common Items

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
B_ABIL	1149	0.1247	0.8718	143.3110	-3.0090	2.3830
FB_ABIL	1149	0.1742	0.9062	200.1360	-2.9870	2.4640

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 1149

	B_ABIL	FB_ABIL
B_ABIL	1.00000	0.99956
	0.0	0.0001
FB_ABIL	0.99956	1.00000
	0.0001	0.0

APPENDIX 5: For Unique Items Only--
Correlation between the "true scores" and various true score estimates, for the three reduced item samples

<Note>
TSCORE = True score
Estimated True Score yielded by Two-stage Equating for Unique Items Only
USIMP12 = Estimates for the simple random sample with 12 common items
USIMP20 = Estimates for the simple random sample with 20 common items
USIMP30 = Estimates for the simple random sample with 30 common items
Estimated True Score yielded by Fixed-b equating for Unique Items Only
UFSIMP12 = Estimates for the simple random sample with 12 common items
UFSIMP20 = Estimates for the simple random sample with 20 common items
UFSIMP30 = Estimates for the simple random sample with 30 common items

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
TSCORE	2241	106.6176	13.4799	238930	44.0000	136.0000
USIMP12	2241	33.1168	3.2176	74215	22.1991	40.5889
USIMP20	2241	27.3623	2.8840	61319	18.3461	34.2786
USIMP30	2241	19.7641	2.1050	44291	13.7153	25.0766
UFSIMP12	2241	33.2057	3.2823	74414	22.2598	40.7793
UFSIMP20	2241	27.3424	2.9387	61274	18.3461	34.3009
UFSIMP30	2241	19.6963	2.1169	44139	13.7153	25.0436

Correlation Analysis

/ Pearson Correlation Coefficients / Prob > |R| under Ho: Rho = 0 / N = 2241

	TSCORE	USIMP12	USIMP20	USIMP30	UFSIMP12	UFSIMP20	UFSIMP30
TSCORE	1.0000	0.8318	0.8470	0.8537	0.8317	0.8467	0.8522
	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
USIMP12	0.8318	1.0000	0.9773	0.9265	0.9993	0.9768	0.9253
	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001
USIMP20	0.8470	0.9773	1.0000	0.9465	0.9779	0.9996	0.9446
	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001
USIMP30	0.8537	0.9265	0.9465	1.0000	0.9260	0.9463	0.9994
	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001
UFSIMP12	0.8317	0.9993	0.9779	0.9260	1.0000	0.9778	0.9243
	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001
UFSIMP20	0.8467	0.9768	0.9996	0.9463	0.9778	1.0000	0.9449
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001
UFSIMP30	0.8522	0.9253	0.9446	0.9994	0.9243	0.9449	1.0000
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000

APPENDIX 6: For Common Items Only– Correlation between the "true scores" and various true score estimates, for the three reduced item samples

```
<Note>
 TSCORE    = True score
 Estimated True Score yielded by Two-stage Equating for Common Items Only
  CSIMP12    = Estimates for the simple random sample with 12 common items
  CSIMP20    = Estimates for the simple random sample with 20 common items
  CSIMP30    = Estimates for the simple random sample with 30 common items
 Estimated True Score yielded by Fixed-b equating for Common Items Only
  CFSIMP12   = Estimates for the simple random sample with 12 common items
  CFSIMP20   = Estimates for the simple random sample with 20 common items
  CFSIMP30   = Estimates for the simple random sample with 30 common items
```

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
TSCORE	2241	106.6176	13.4799	238930	44.0000	136.0000
CSIMP12	2241	8.0957	1.1063	18142	4.6514	10.4574
CSIMP20	2241	13.8994	1.5741	31149	8.9312	17.4499
CSIMP30	2241	21.8360	2.0762	48934	15.1866	26.4787
CFSIMP12	2241	8.1256	1.1265	18209	4.6663	10.5060
CFSIMP20	2241	13.8877	1.6041	31122	8.9312	17.4601
CFSIMP30	2241	21.7685	2.0946	48783	15.1866	26.4546

Correlation Analysis

/Pearson Correlation Coefficients /Prob>|R| under Ho: Rho=0 /N=2241

	TSCORE	CSIMP12	CSIMP20	CSIMP30	CFSIMP12	CFSIMP20	CFSIMP30
TSCORE	1.0000	0.8315	0.8474	0.8568	0.8315	0.8471	0.8552
	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
CSIMP12	0.8315	1.0000	0.9773	0.9267	0.9993	0.9769	0.9255
	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001
CSIMP20	0.8474	0.9773	1.0000	0.9468	0.9778	0.9996	0.9449
	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001
CSIMP30	0.8568	0.9267	0.9468	1.0000	0.9264	0.9468	0.9994
	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001
CFSIMP12	0.8315	0.9993	0.9778	0.9264	1.0000	0.9778	0.9247
	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001
CFSIMP20	0.8471	0.9769	0.9996	0.9468	0.9778	1.0000	0.9454
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001
CFSIMP30	0.8552	0.9255	0.9449	0.9994	0.9247	0.9454	1.0000
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000